REFINED GENETIC ALGORITHMS
FOR
POLYPEPTIDE STRUCTURE PREDICTION

REFINED GENETIC ALGORITHMS

FOR

POLYPEPTIDE STRUCTURE PREDICTION

Acknowledgements

Table of Contents

List of Figures

List of Tables

List of Symbols

α

$\mathcal{F}$

$\mathcal{S}$

α

γ

i i

ϕ

ψ

ω

χ_i i

List of Abbreviations

Abstract

80

REFINED GENETIC ALGORITHMS
FOR
POLYPEPTIDE STRUCTURE PREDICTION

I. Introduction

1.1 Protein Folding Problem / Polypeptide Structure Prediction

predict

native conformation[1]

$$PSP \subseteq PFP$$

[1]The *native conformation* determines the protein's biological functions.

1.1.1 Background.

α

polypeptide *peptide* *protein*

2

residue

primary structure

secondary structure

tertiary structure

1.1.2 Importance.

Grand Challenges

1.1.3 Methods for Polypeptide Structures Prediction.

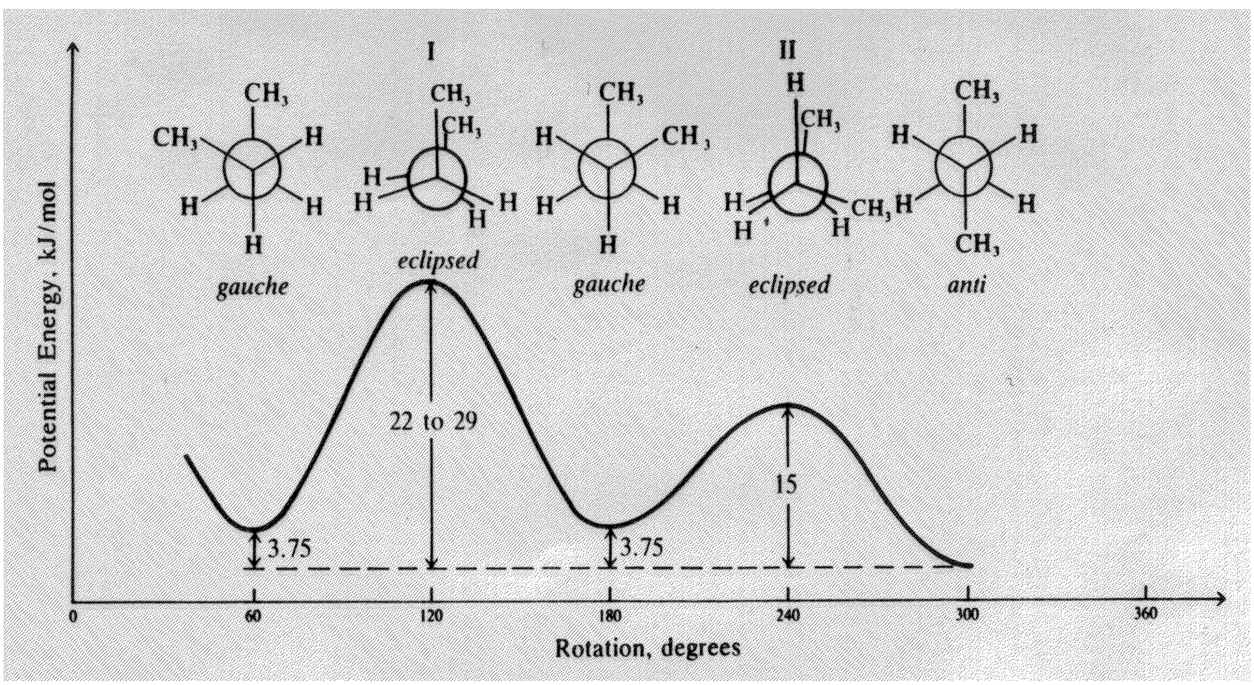
energy minimization *molecular dynamics*

ab initio *semi-empirical*

force-field

$\mathcal{O}$ n^5 $\mathcal{O}$ n^4 $\mathcal{O}$ n^2

n



n

-15

-12

1.1.4 Growth of Complexity.

x, y z

$| n- |$

n [2]

$|3n-6|$

1.2 Genetic Algorithms

implicit parallelism[3]

Evolutionary Computing

chromosome[4]

[2] Regardless of whether a *Cartesian* or *Internal* coordinate system is used. However, the internal coordinate system has fewer *independent variables*

[3] See Appendix B.3.3

[4] Because GAs are loosely based on natural evolution, many of the terms associated with natural evolution are used interchangeably with the terms created specifically for genetic algorithms (67).

[5]

$< selection,\ crossover,\ mutation >$

generation *genes*

Alleles

YES

1.3 Parallel and Distributed Computing

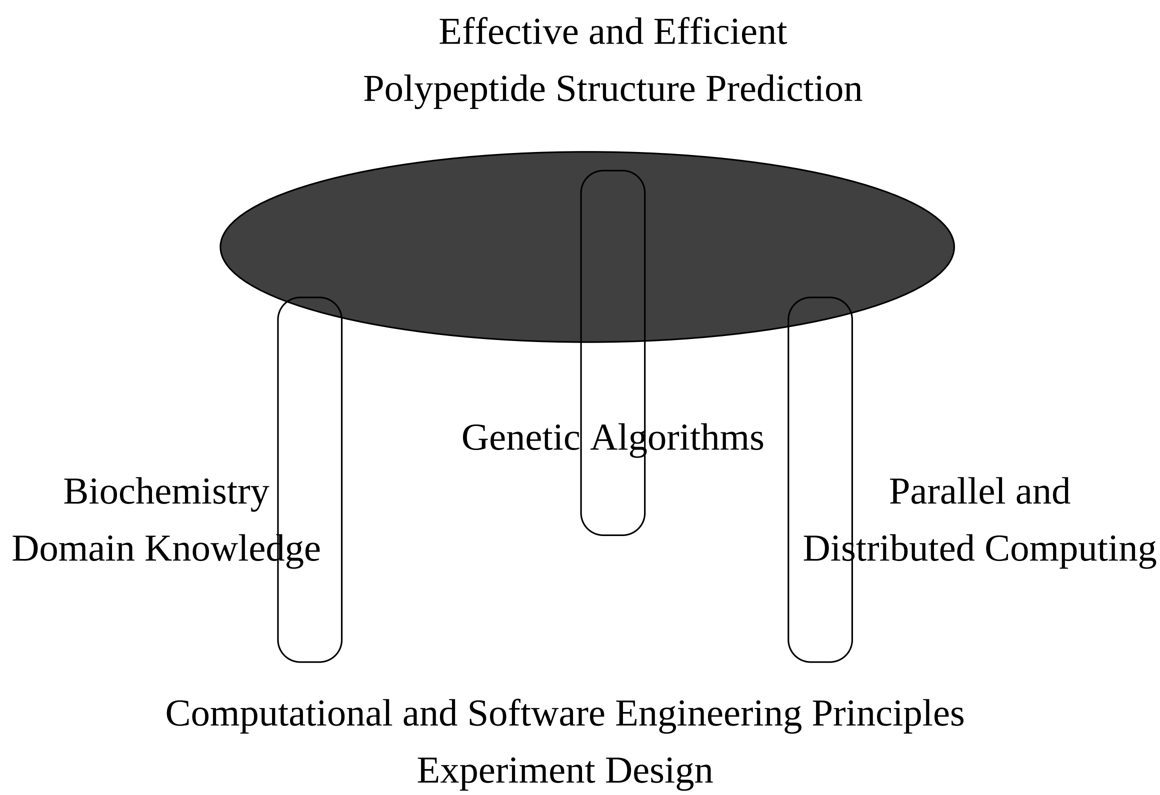
single instruction stream, multiple data stream

multiple instruction stream, multiple data stream

scalable

1.4 Research Objectives

[5] Crossover is sometimes called recombination.



- *Improve Performance of Hybrid GAs for PSP*

- *Real Valued Genetic Algorithm Implementation for the PSP*

- *Exploit Domain Knowledge to Limit Search Space*
  domain knowledge

1.5 *Methodology*

Evolution Program[6] *exogenous* [7] [8]

1.6 *Assumptions*

[9]

1.7 *Summary*

---

[6] Not to be confused with *Evolutionary Programming*, see Section 2.4.

[7] The probability that an improvement at a specific node is migrated to other nodes.

[8] The probability, given a migration, that it is migrated to all other nodes.

[9] Or molecular conformation

II. Current Issues

2.1 Introduction

2.2 Previous Research

2.3 Polypeptides Structure Prediction (PSP)


Genetic Algorithm
Parallel Implementation (Sawyer)
Communication Strategies (Merkle)
Protein Folding Problem Initial Energy Model (Brinkman)
Protein Folding Problem Energy Model Refinement (Gates)
Protein Folding Problem Serial Hybrid GA (Gaulke)
Premature Convergence Strategies (Dymek)
Messy Genetic Algorithm
Initial Implementation GA-hard Problems (Dymek)
Combinatoric Optimization Data Distribution Strategies (Merkle)
Fast Messy Genetic Algorithm
Initial Implementation Parallel Implementation (Merkle)
Additional Parallelization Protein Folding Problem (Gates)
Permutation Genetic Algorithm

$\alpha\beta\beta$

$\epsilon - optimal$

$\alpha\beta\beta$

ω $trans^1$ ϕ, ψ

ϕ, ψ, ω

$\chi's$

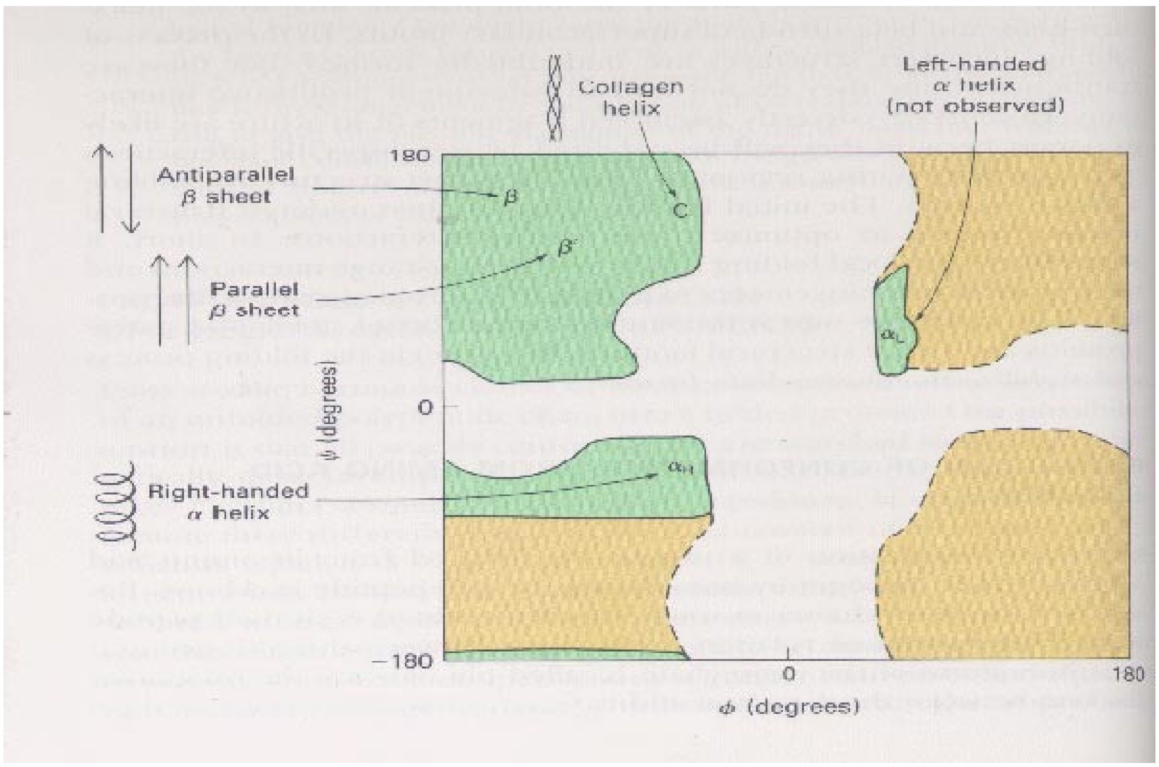
ϕ, ψ

---

1 $\pm 180^{circle}$ or $\pm \pi$

Ramachandran Plot ϕ

ψ



Biochemistry

2.4 Genetic Algorithms

evolution program

paradox of real codings

virtual alphabets

blocked

Fundamental Theorem of Genetic Algorithms

strong *weak*

EP_1

EP_5 EP_1 EP_5

P

Q

P EP_1 EP_5

2.5 Parallel Genetic Algorithms

is

data parallelism

island model *neighborhood model*

farming model

Island Model:

migration

$\mathcal{O}\ \frac{nl}{p}$ p n

$p \ll n$ $p \rightarrow n$

Neighborhood Model:

n p

$\mathcal{O}\ s\ \ l$ $\mathcal{O}\ nl$ s

Farming Model: *farming*

farm out

foreman *workers*

2.6 Summary

III. Algorithm Analysis, Design, and Implementation

3.1 Analysis

3.1.1 Cost Analysis of Local Minimization using Conjugate Gradient.

Numerical Recipes in C

n n

$\mathcal{O}$ n^2 `ITMAX = 200`

Lamarckian *Baldwinian*

`eval_func()`

`ITMAX`

generational

checkpoint *asynchronous*

farming

client

server

a priori

		_	_	_	

`charm_eval()`

`charm_eval()`

`charm_eval()`

`charm_eval()`

3.1.2 Constraint Set Development.

3.1.2.1 Conventions Adopted.

θ_{min} θ_{max} θ_{min} θ_{max}

θ_{min} θ_{max}

$$\theta_{min} > \theta_{max} \rightarrow \theta_{min_{adj}} \quad \theta_{min} - \quad ^{\circ}$$

$$\theta_{min} < \theta_{max} \rightarrow \theta_{max_{adj}} \quad \theta_{max} \quad ^{\circ}$$

$\theta_{min_{adj}}$ $\theta_{max_{adj}}$

θ_{min} θ_{max}

3.1.2.2 [Met]-enkephalin.

phi *psi* $\phi\psi$

3.1.2.3 Polyalanine.

a priori α $\phi\psi$ α

			θ_{min}	θ_{max}
$Non-glycine$	$-$		$-$	$-$
$Glycine$	$-$		$-$	$-$
			$-$	$-$
	$-$		$-$	$-$
χ_1	$- \mid \mid$		$- \mid \mid -$	$- \mid \mid -$

			θ_{min}	θ_{max}
$Non-glycine$	$-$		$-$	
$Glycine$			$-$	$-$
			$-$	$-$
		.	. $-$	$-$.
χ_1	$- \mid \mid$	.	$-$. $\mid$. $\mid$. $-$	$-$. $\mid$. $\mid -$.

α

			θ_{min}	θ_{max}
	$-$.	.	$-$	$-$
	$-$		$-$	
			$-$	$-$
χ_1	$- \mid \mid$		$- \mid \mid -$	$- \mid \mid -$

3.1.3 Real-valued GAs.

				θ_{min}	θ_{max}
		$-$		$-$	$-$
		$-$		$-$	$-$
				$-$	$-$
χ_1	$-$ | |		$-$ | | $-$	$-$ |	

[1]

3.2 Algorithm Design and Implementation

3.2.1 Parallel Hybrid GA.

farming model

3.2.1.1 Algorithm Design.

G_{t+1} G_t

[1]The developers of this algorithm have been inconsistent with its naming in the literature. Genocop, Genocop-III.1.0, GENOCOP-III, etc., have been used synonymously. In this document I have adopted the standardization of GENOCOP-III.

Q

$EvalCnt \leftarrow population\ size$

Loop

Until $EvalCnt$

Loop

$EvalCnt \leftarrow$

Loop

Until $EvalCnt$

Until

3.2.1.2 Scheduling. *round robin*

network of workstations

3.2.2 REal-valved GA, Limited by constraints (REGAL) .

evolution program

Loop

Until

$REPLACE$

$server_i$ Q

$EvalsPerformed$

$EvalsPerformed < EvalCnt \wedge Q \neq \emptyset$

$server_j$ Q

$server_j$

3.2.2.1 Incorporation of Domain Knowledge.

Evolution Program

probable *improbable* $\mathcal{S}_{prob}$

$\mathcal{S}_{improb}$

$\mathcal{S}_{prob}$ $\mathcal{S}_{improb}$ $\mathcal{S}$

$\mathcal{S}_{prob}$ $\mathcal{S}_{improb}$ $\emptyset$

$\theta_{i_{min}}$ $\theta_{i_{max}}$ $x_i \in \vec{x} \in \boldsymbol{R}$

x_i $\theta_{i_{min}}$ $\theta_{i_{max}}$ $-\pi$ π

$$\leq \quad \theta - \frac{\theta_{min} \quad \theta_{max}}{} \quad - \quad \frac{\theta_{max} - \theta_{min}}{}$$

$\{\phi, \psi, \omega\}$

$$\leq \quad \theta - \frac{\theta_{min} \quad \theta_{max}}{} \quad - \quad \frac{\theta_{max} - \theta_{min}}{}$$

$\{\chi_1\}$

3.2.3 Parallel REGAL (Para-REGAL). *modified island*

island

[2]

probabilistic migration *Probability of Migration*

P_m *Probability of Complete Migration* P_{cm}

P_m P_{cm}

i j

$P_m \cdot P_{cm}$

$P_m \cdot P_{cm} >$.

archipelago [3]

- *Average Genotypic Distance*

- *Least Genotypic Distance*

---

[2] Except for the seed of the random number generator.

[3] A collection of islands; thus the islands making of the Para-REGAL execution

- *Greatest Genotypic Distance*
- *Average Best Fitness*
- *Local Delta*

3.3 Summary

IV. Experiment Design

4.1 Experiment Techniques

4.1.1 Random Number Seeds.

nodal

```
Seed =(unsigned) ((Seed + My_node) / (My_node + 1);
```

```
Seed =(unsigned) ((Seed + My_node) % Max_Seed_Value);
```

$$Seed \quad floor \quad * rand\ number\ from\ Xcalculator$$

$$Seed \quad floor \quad * randon\ number\ from\ Xcalculator$$

$$Seed \quad floor \quad * randon\ number\ from\ Xcalculator$$

4.1.2 Statistical Techniques.

4.1.2.1 Analysis of Variance (ANOVA).

a

n a

σ^2

4.1.2.2 Kruskal-Wallis H Test.

4.2 Experiment I: Evaluation of the Efficiency of a Parallel & Distributed Hybrid GA

4.2.1 Motivation and Objective.

4.2.2 Methodology.

P_m

P_r

farming model

			P_m	P_r

987654321

4.3 Experiment II: Evaluation of the Use of Constraints in the PSP

4.3.1 Motivation and Objective.

4.3.2 Methodology.

none *loose* *tight*

[Met]-enkephalin tight *Polyalanine tight*

4.3.3 Parameter Selection.

reference population

$\mathcal{F}$ $\mathcal{S}$

[1]

loose *tight*

4.4 Experiment III: Evaluation of Exogenous Parameters in the REGAL System

4.4.1 Motivation and Objective.

[1]Based on the ratio of $\frac{\mathcal{F}}{S}$, it would require 10^{67} tries to randomly generate just on fully feasible chromosome when using the *tight* constraint set for Polyalanine.

4.4.2 Methodology.

iterations

steady state

adaptive

4.4.3 Exogenous Parameter Evaluation Experiments.

Offsprings *Reference Population Size*

	H_0 H_1

4.5 Experiment IV: Evaluation of Para-REGAL

4.5.1 Motivation and Objective.

4.5.2 Methodology.

Island 0

Islands 1 & 2 *loose* *Island 3* *tight*

,

4.5.3 Para-REGAL Experiments.

$\{ \ . \ , \ . \ , \ . \ , \ . \ \}$ P_m P_{cm} ,

		ω		

4.6 Summary

V. Results and Analysis

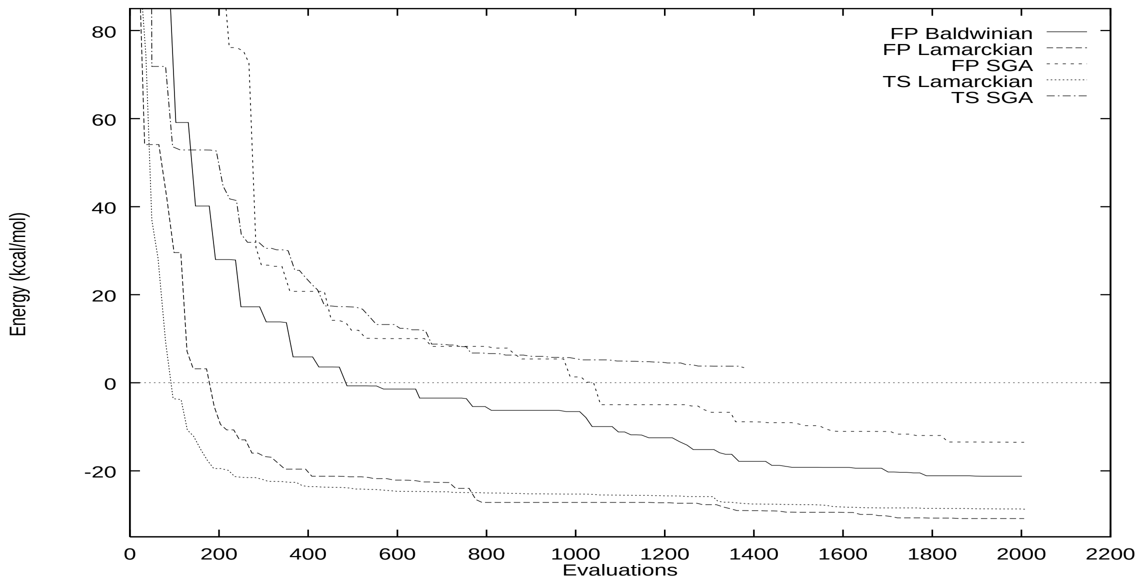
5.1 Experiment I: Parallel Hybrid GA

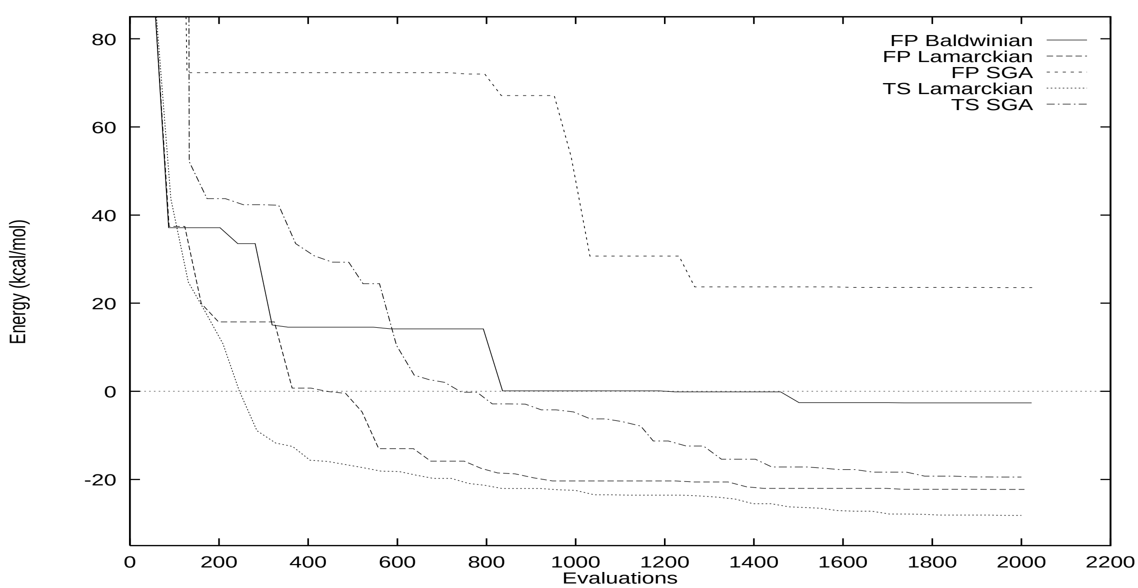
5.1.1 Effectiveness Analysis.

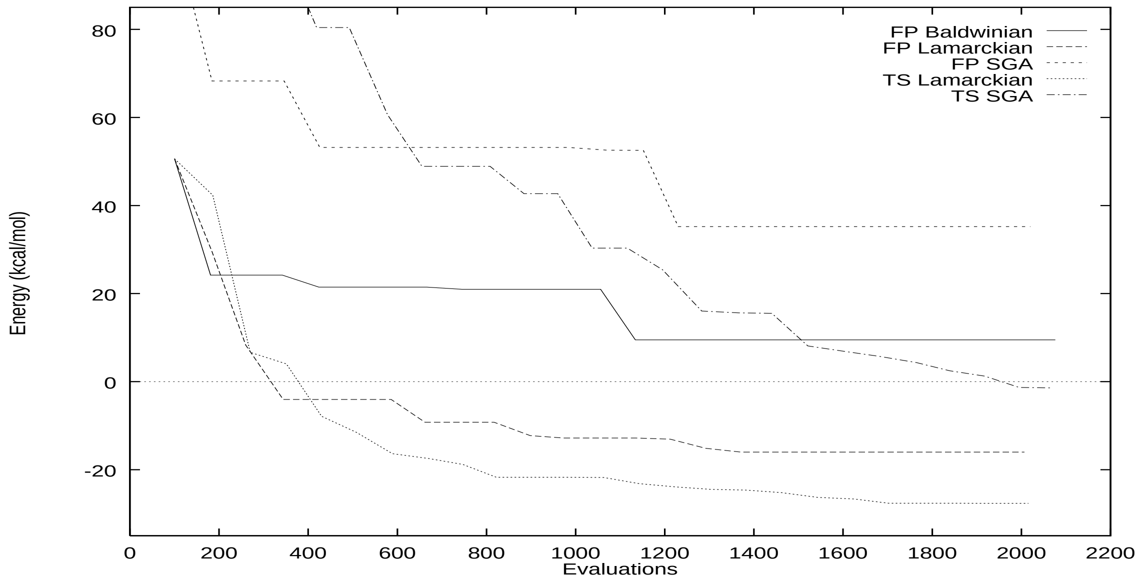
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

selection pressure

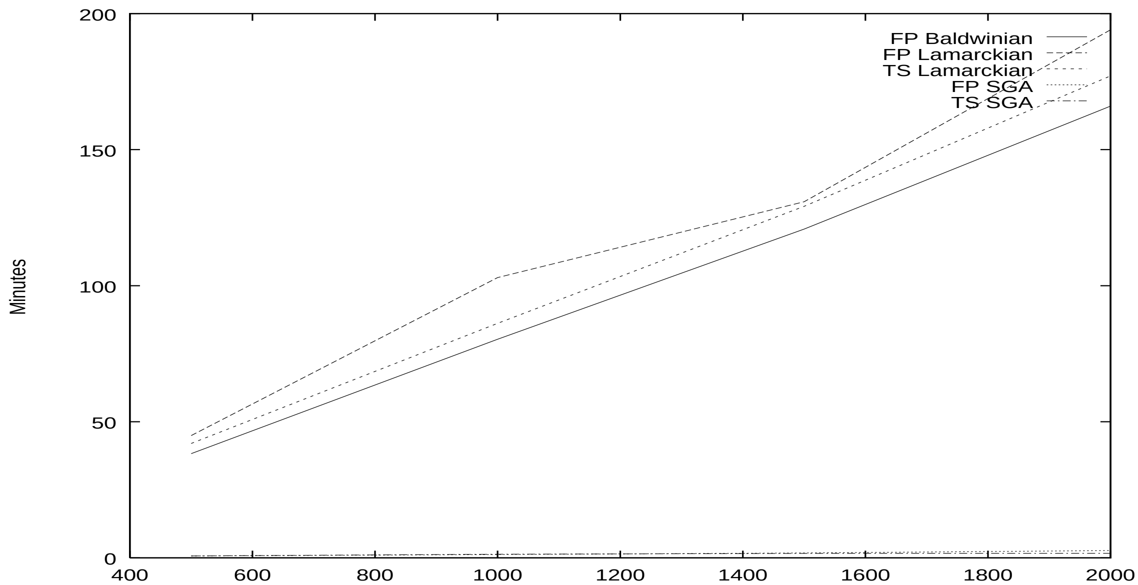
x $\frac{x}{pop\_size_{small}} > \frac{x}{pop\_size_{large}}$

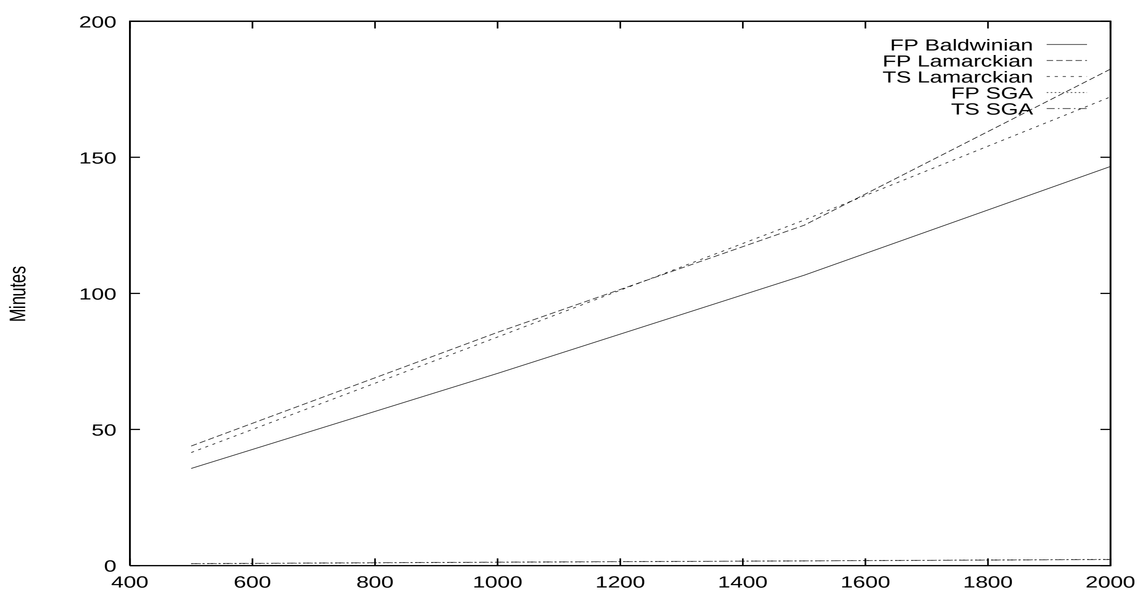

FP Baldwinian
FP Lamarckian
FP SGA
TS Lamarckian
TS SGA
Energy (kcal/mol)
Evaluations

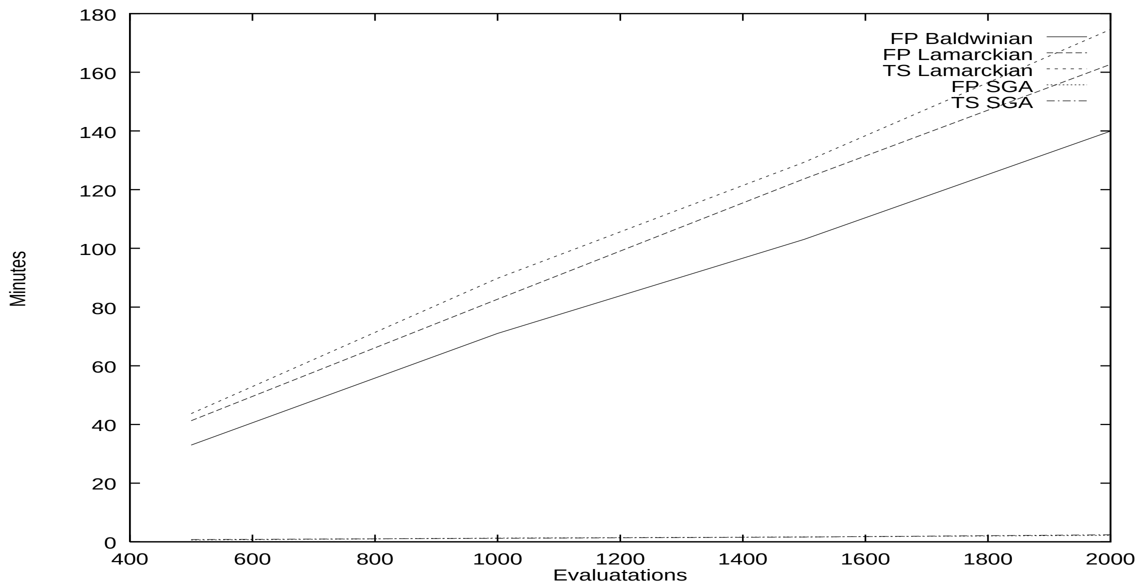

FP Baldwinian
FP Lamarckian
FP SGA
TS Lamarckian
TS SGA
Energy (kcal/mol)
Evaluations

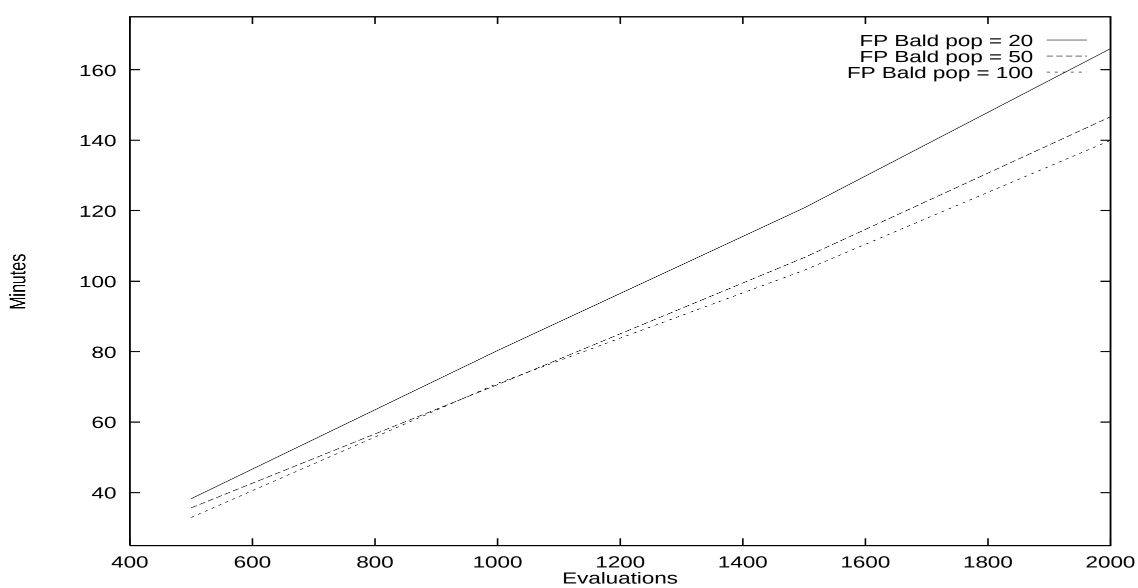


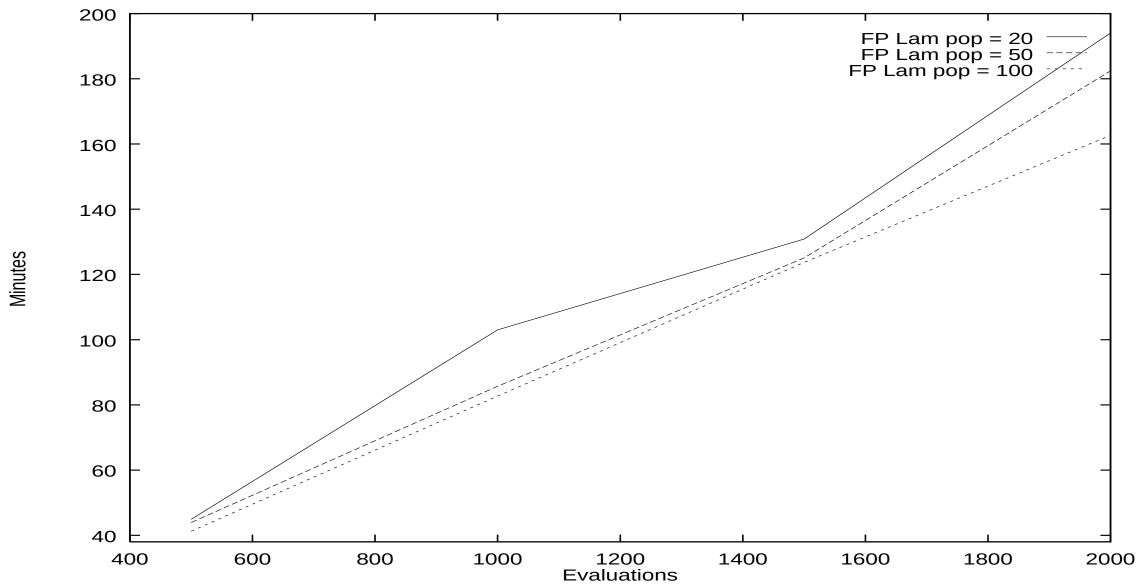
5.1.2 Efficiency Analysis, Serial.


Minutes
200
150
100
50
0
400
600
800
1000
1200
1400
1600
1800
2000
FP Baldwinian
FP Lamarckian
TS Lamarckian
FP SGA
TS SGA

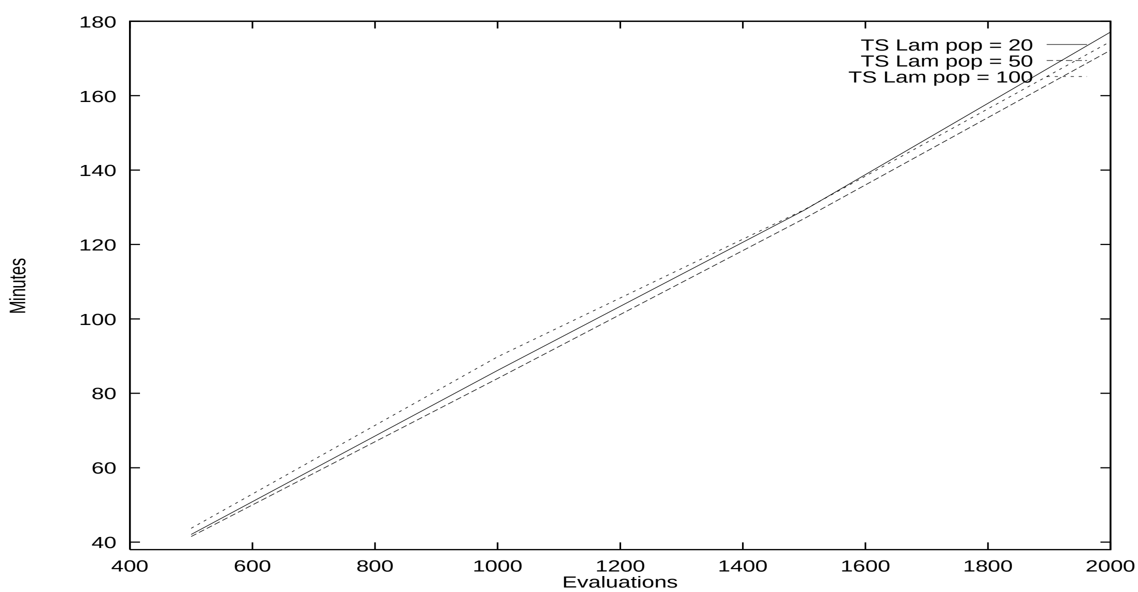

Minutes
200
150
100
50
0
400
600
800
1000
1200
1400
1600
1800
2000
FP Baldwinian
FP Lamarckian
TS Lamarckian
FP SGA
TS SGA


FP Baldwinian
FP Lamarckian
TS Lamarckian
FP SGA
TS SGA
Minutes
Evaluatations
0
20
40
60
80
100
120
140
160
180
400
600
800
1000
1200
1400
1600
1800
2000

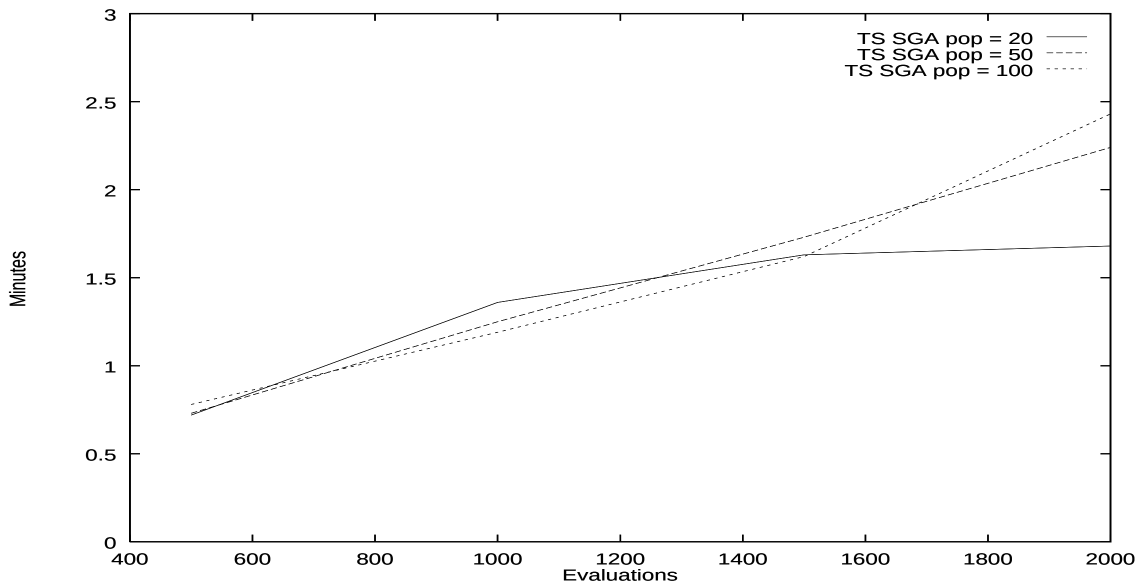

FP Bald pop = 20
FP Bald pop = 50
FP Bald pop = 100
Minutes
Evaluations
40
60
80
100
120
140
160
400
600
800
1000
1200
1400
1600
1800
2000



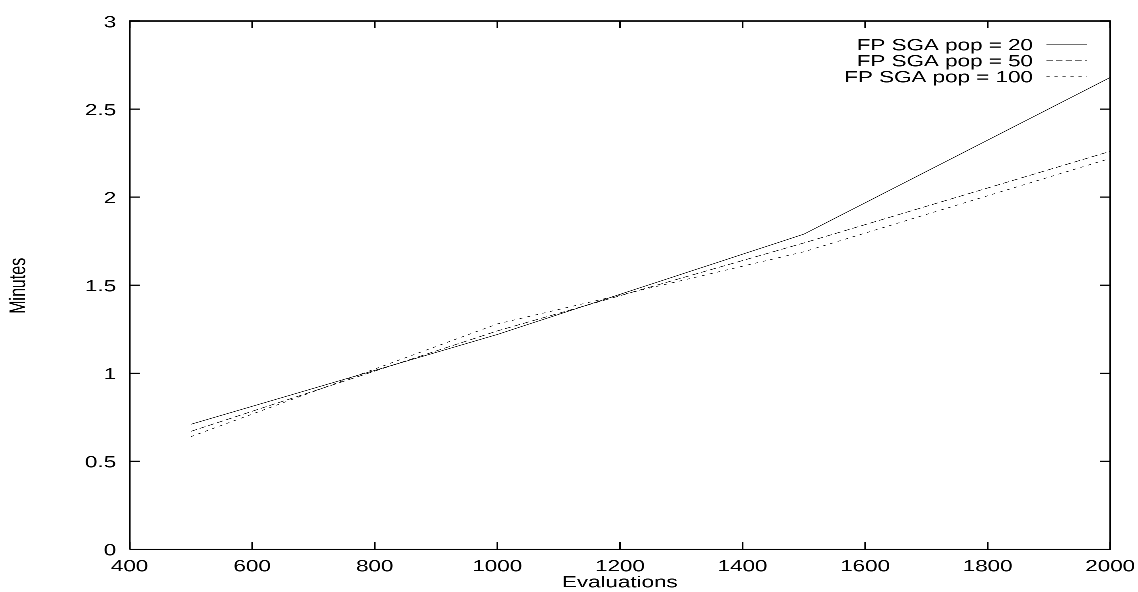
Minutes
Evaluations
FP Lam pop = 20
FP Lam pop = 50
FP Lam pop = 100



Minutes
Evaluations
TS Lam pop = 20
TS Lam pop = 50
TS Lam pop = 100



TS SGA pop = 20
TS SGA pop = 50
TS SGA pop = 100
Minutes
Evaluations
0
0.5
1
1.5
2
2.5
3
400
600
800
1000
1200
1400
1600
1800
2000



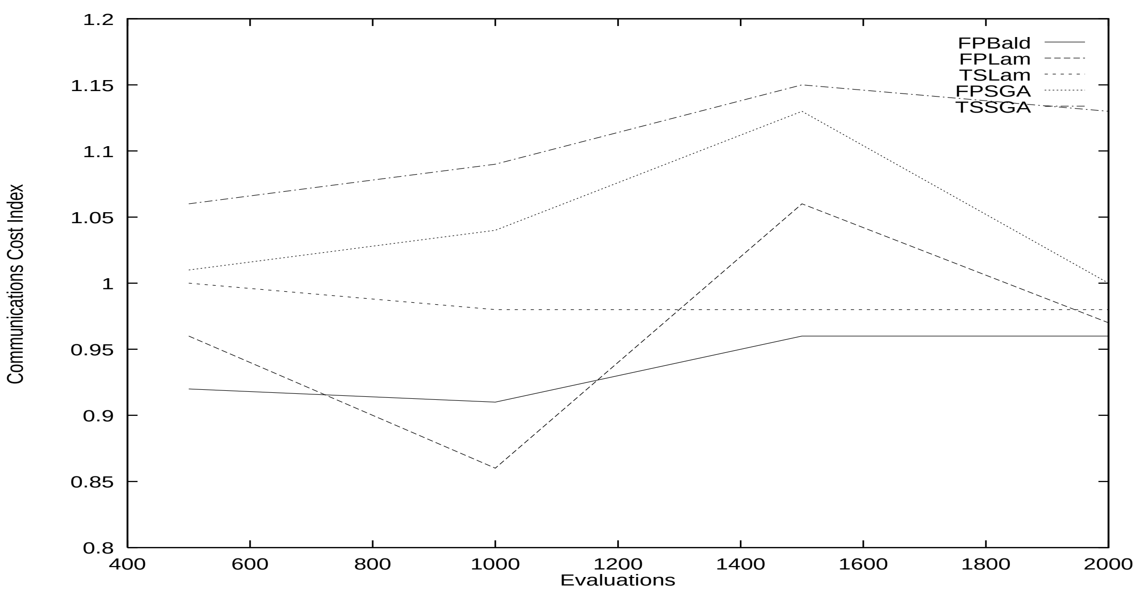
FP SGA pop = 20
FP SGA pop = 50
FP SGA pop = 100
Minutes
Evaluations
0
0.5
1
1.5
2
2.5
3
400
600
800
1000
1200
1400
1600
1800
2000

5.1.3 Efficiency Analysis, Parallel.

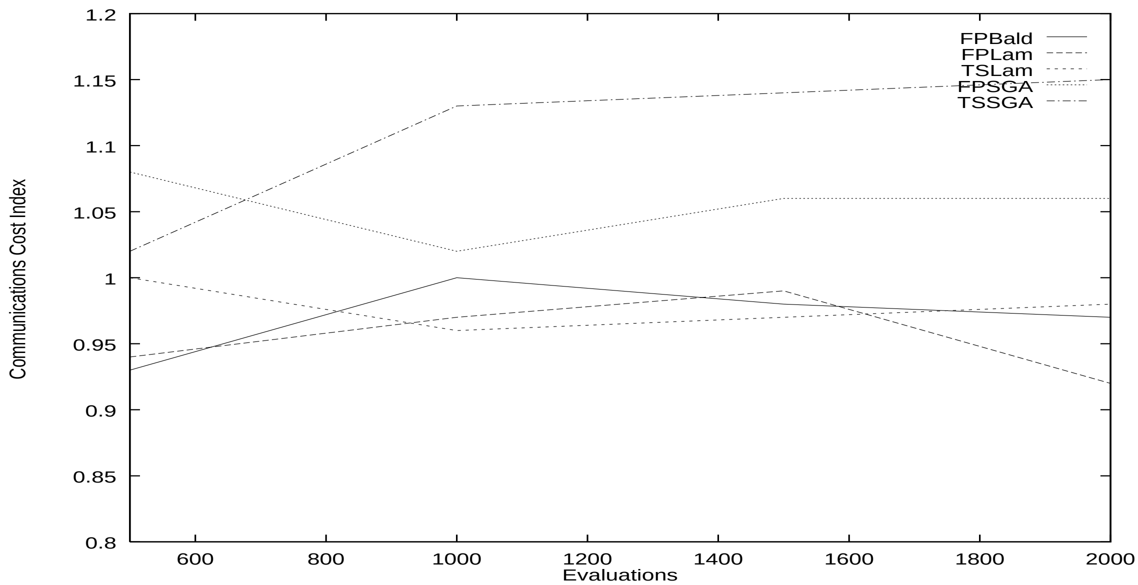
$$C_{communications} \quad T_{P_2} - T_S$$

$C_{communications}$ T_{P_2} T_S

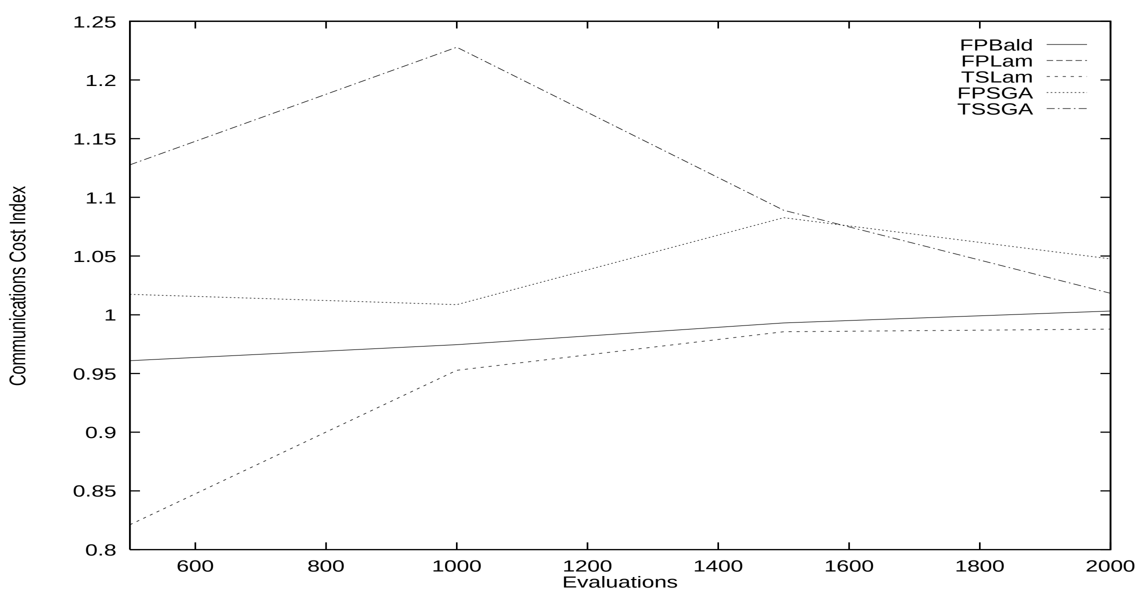
$$Communication\ Cost\ Index \quad \frac{T_{P_2}}{T_S}$$



$$S \quad \frac{T_S}{T_P}$$



Communications Cost Index
Evaluations
FPBald
FPLam
TSLam
FPSGA
TSSGA

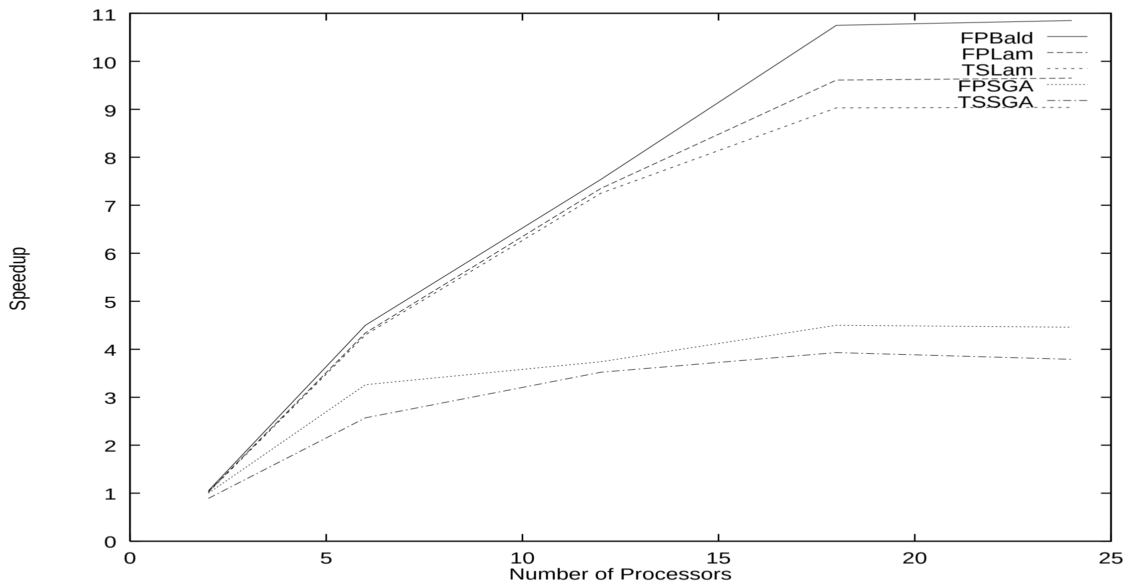


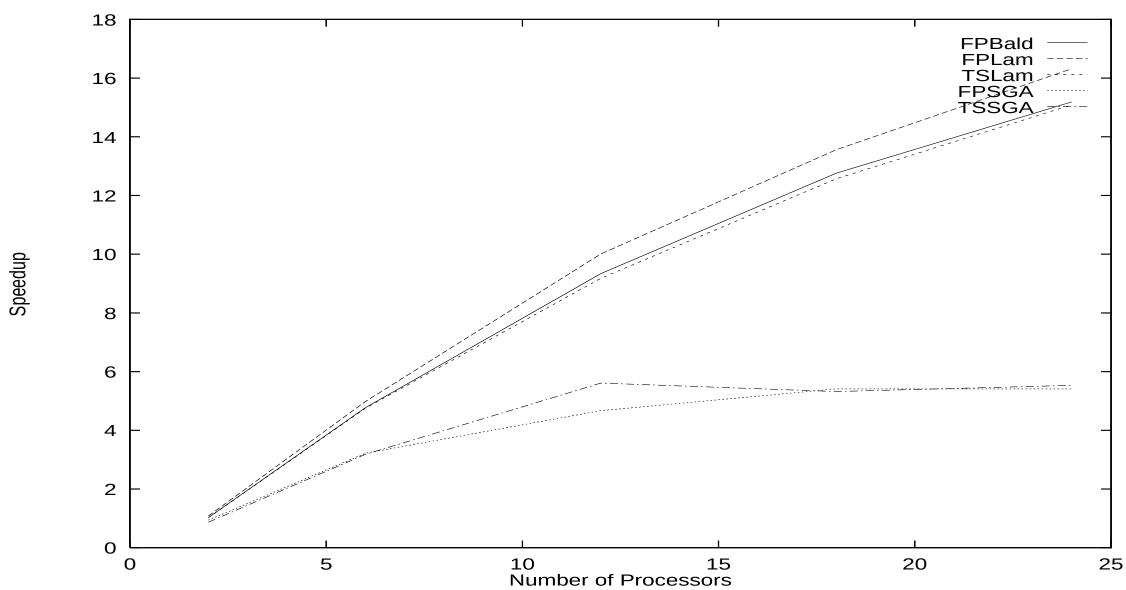
Communications Cost Index
Evaluations
FPBald
FPLam
TSLam
FPSGA
TSSGA

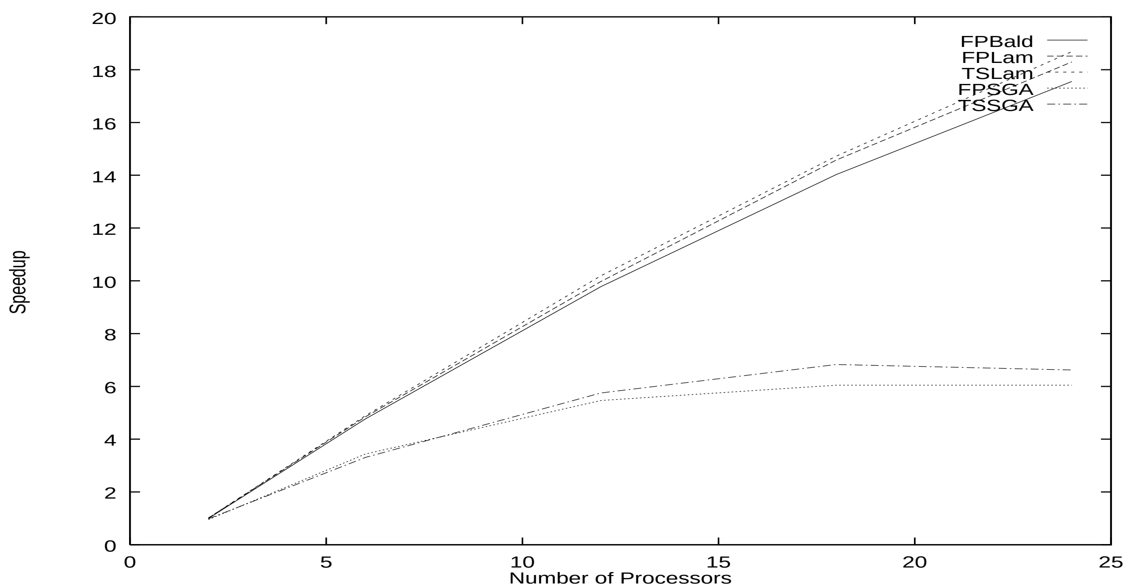
S T_S *best* T_P

linear speedup S

p $S > p$ *super linear speedup* T_S







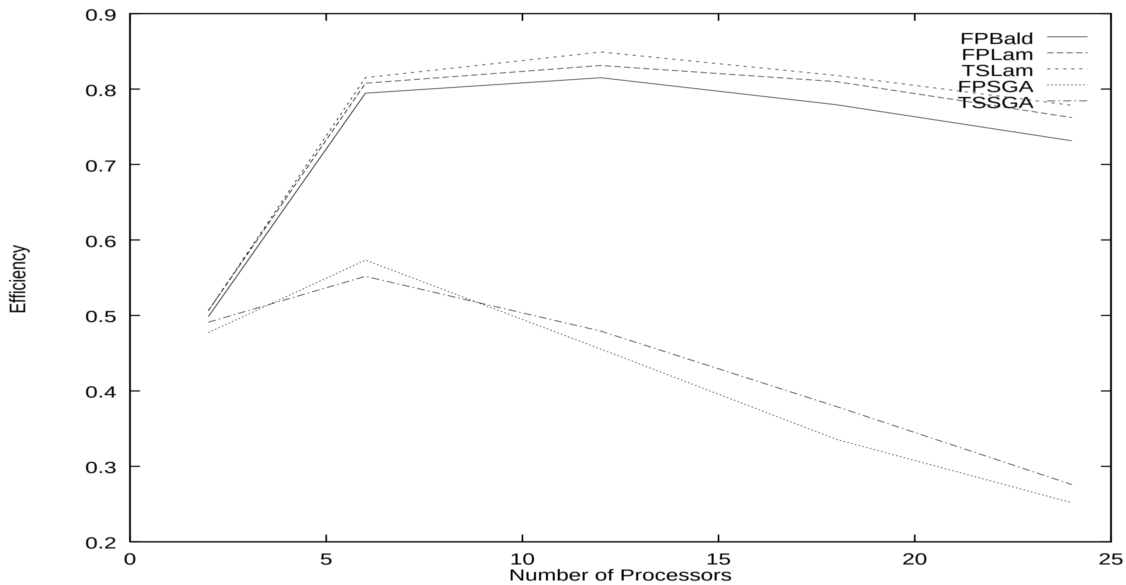
$$E \quad \frac{S}{p}$$

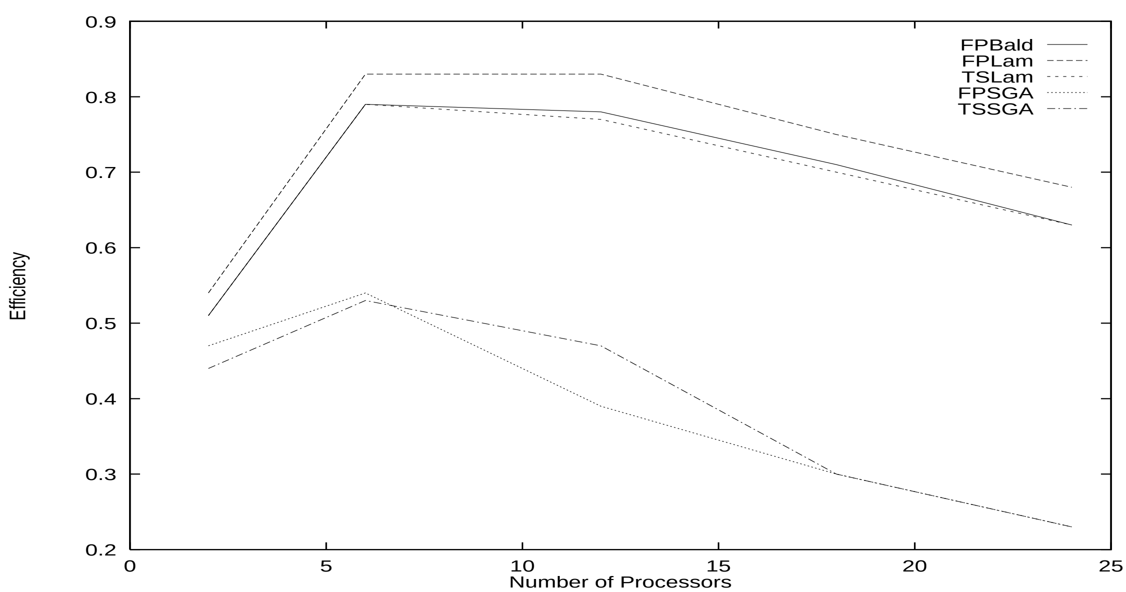
E S p

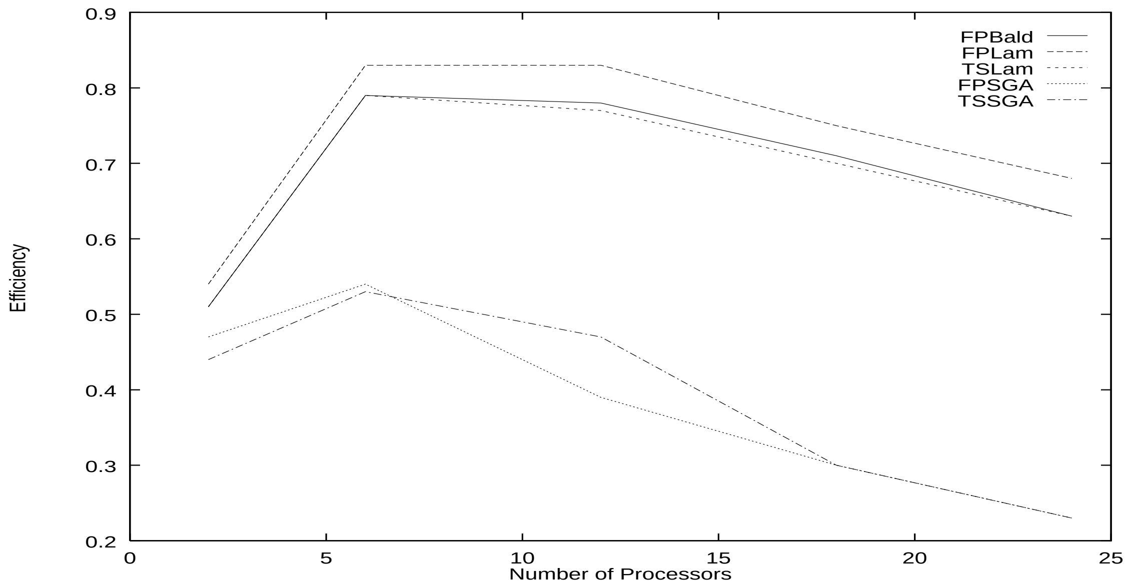
$Cost$

cost-optimal

5.2 Experiment II: Preliminary REGAL Evaluation


FPBald
FPLam
TSLam
FPSGA
TSSGA
Efficiency
Number of Processors


FPBald
FPLam
TSLam
FPSGA
TSSGA
Efficiency
Number of Processors



5.2.1 [Met]-enkephalin.

$\mathcal{F}$ $\mathcal{S}$

5.2.2 *Polyalanine.*

step

α

α

5.2.3 *Efficiency.*

-
  -

		1	

5.3 *Experiment III: Analysis of Exogenous Parameters for REGAL*

independent

[2]

current best trajectory

[3]

[2] Analysis of Variance, see Appendix F.1 for more details. Concern has be raised about lack of variability because a single seed set was used. The Kruskal-Wallis H Test (Appendix F.2) was used as an independent method to verify the ANOVA results. The conclusions were the same. Kruskal-Wallis results are not shown

[3] Hypothesis testing was not done on run times because system loading in the multi-user environment could not be controlled. They are provided for reference only. However, the large number of experiments tends to dampen out cases were the platform was heavily loaded. Thus, the data are insightful.

4 This nomenclature is from the GENOCOP-III documentation. It would be more accurate to stay the reference population is operated upon.

$\{ \cdot , \cdot , \cdot \}$

$\cdot , \cdot$

				F_0	α .
				*	
				*	

α . *

α . H_0

α α

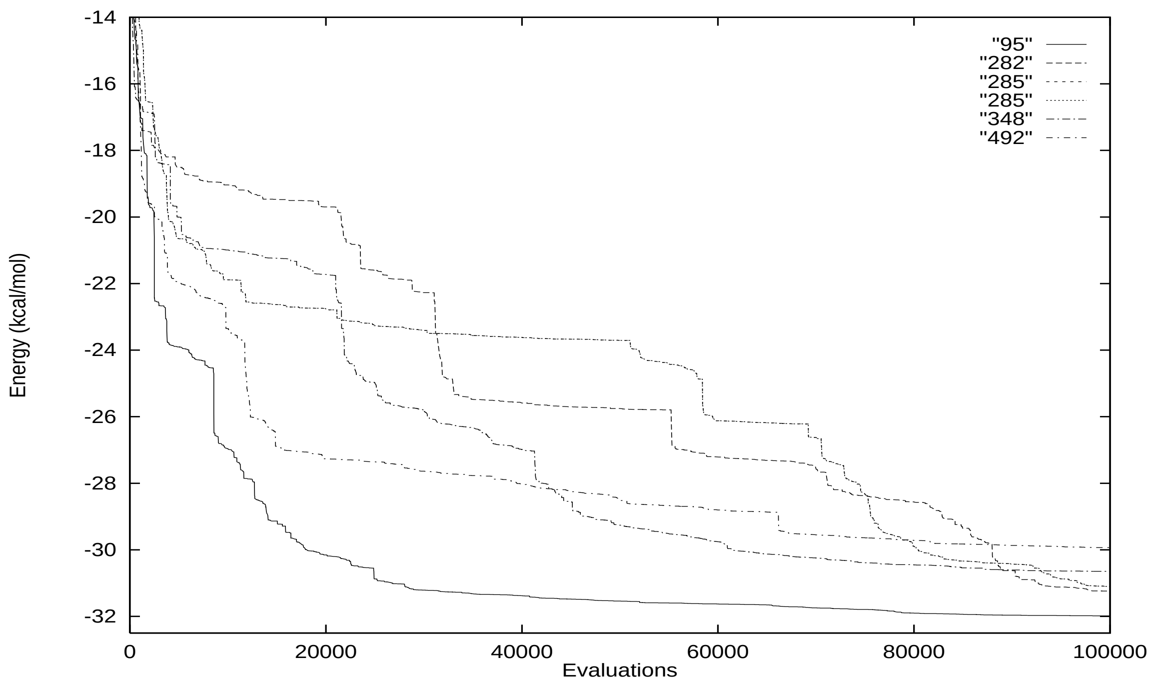
$\cdot , \cdot$ α α^{-i} $\leq i \leq$

α .

5.4 Experiment IV: Analysis of Para-REGAL

P_m P_{cm} .

Island 3



periodicity

m

n $\frac{m}{n}$ m

$^{-3}$ th

$-$ K

	P_m	P_{cm}			

5.5 Summary

[5]

[5] Considerable resources is a relative concept. Even a 1000 hours of computer time is trivial when compared to experimental techniques that require years to yield results.

VI. Conclusions and Recommendation

Grand Challenge

6.1 Initiative I: PHGA

6.2 Initiative II: REGAL

trans [1]

trans [2]

[1] Actual implementation is out of scope for this investigation because research is required into appropriate control metrics.
[2] Limit the dihedral angle's range to a lower bound greater than $-\pi$ and an upper bound less than π.

fundamental theorem of genetic algorithms

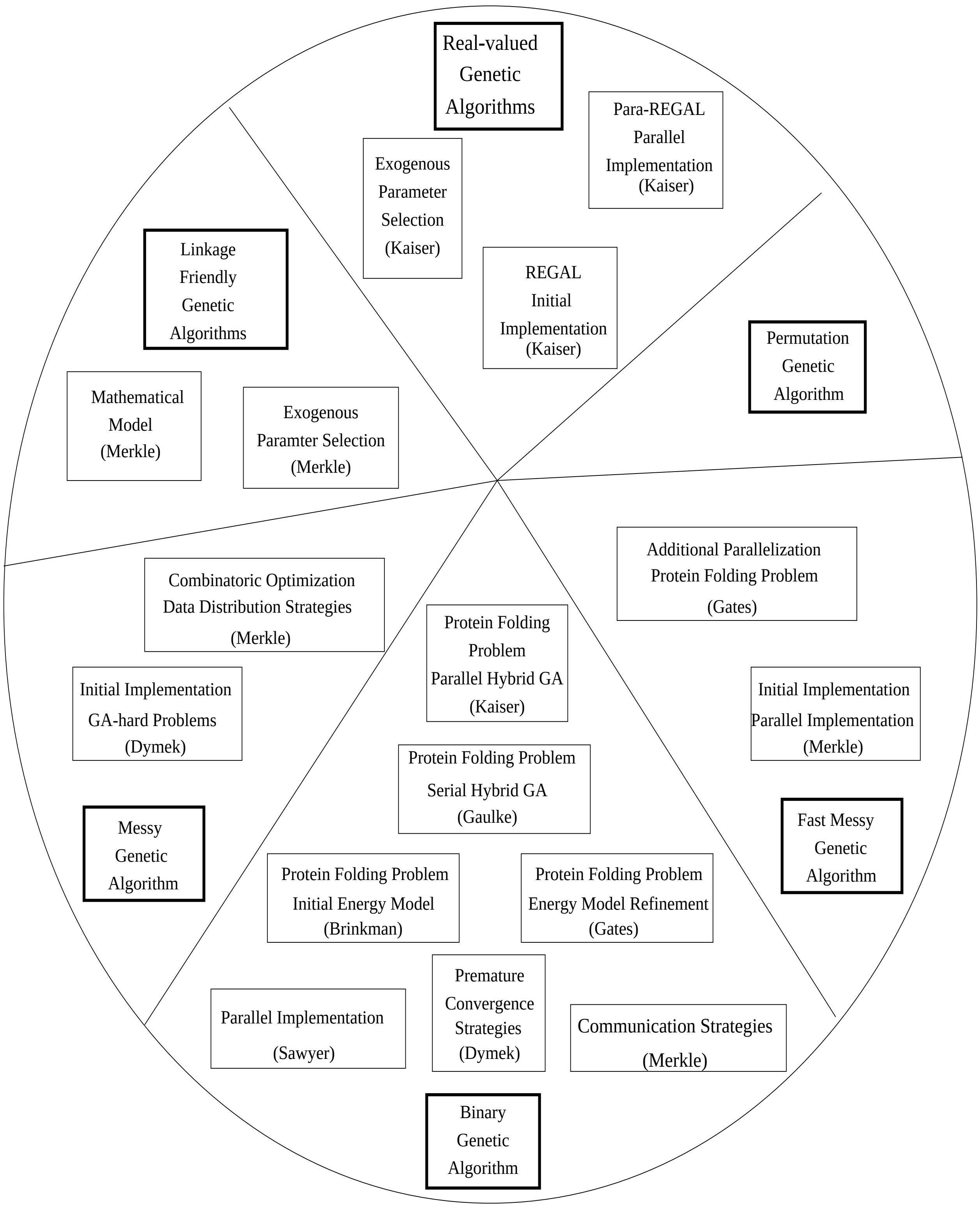
6.3 Initiative III: Examination of Exogenous Parameters

α

6.4 Initiative IV: Para-REGAL

molten globular

α



Real-valued Genetic Algorithms
Para-REGAL Parallel Implementation (Kaiser)
Exogenous Parameter Selection (Kaiser)
REGAL Initial Implementation (Kaiser)
Linkage Friendly Genetic Algorithms
Permutation Genetic Algorithm
Mathematical Model (Merkle)
Exogenous Paramter Selection (Merkle)
Additional Parallelization Protein Folding Problem (Gates)
Combinatoric Optimization Data Distribution Strategies (Merkle)
Protein Folding Problem Parallel Hybrid GA (Kaiser)
Initial Implementation GA-hard Problems (Dymek)
Initial Implementation Parallel Implementation (Merkle)
Protein Folding Problem Serial Hybrid GA (Gaulke)
Messy Genetic Algorithm
Fast Messy Genetic Algorithm
Protein Folding Problem Initial Energy Model (Brinkman)
Protein Folding Problem Energy Model Refinement (Gates)
Parallel Implementation (Sawyer)
Premature Convergence Strategies (Dymek)
Communication Strategies (Merkle)
Binary Genetic Algorithm

6.5 Recommendations

6.6 Summary

Appendix A. Background on the Protein Folding and Protein Structure Prediction Problems

inverse folding problem

evaluation

design

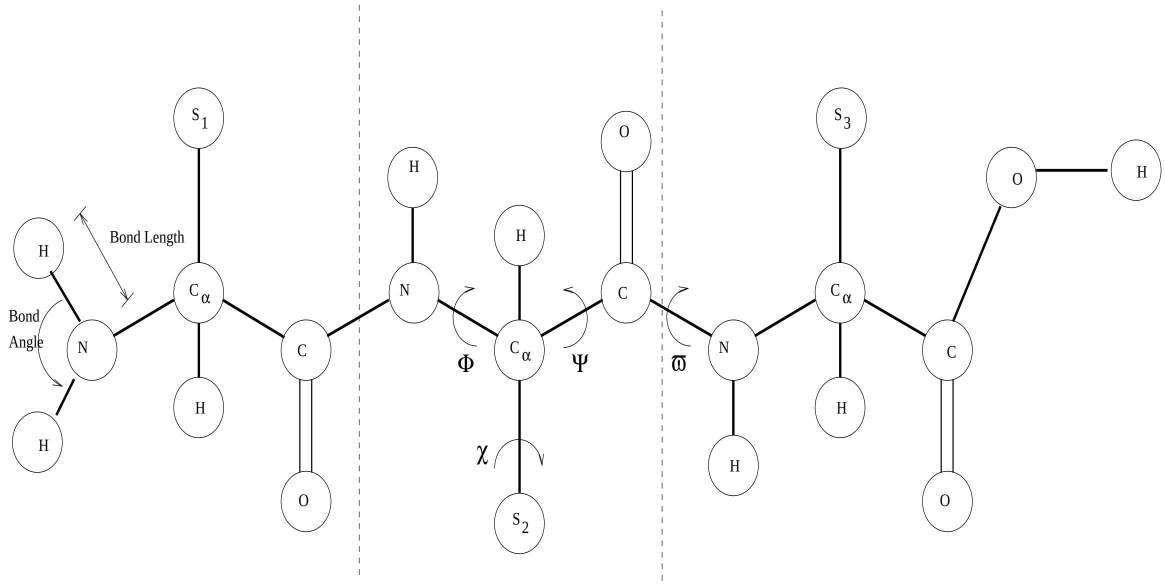
A.1 Introduction to Proteins and Associated Terminology

backbone $-$ α γ

side-chain i α

primary structure

residues



α-helices
β-sheets *secondary structure*

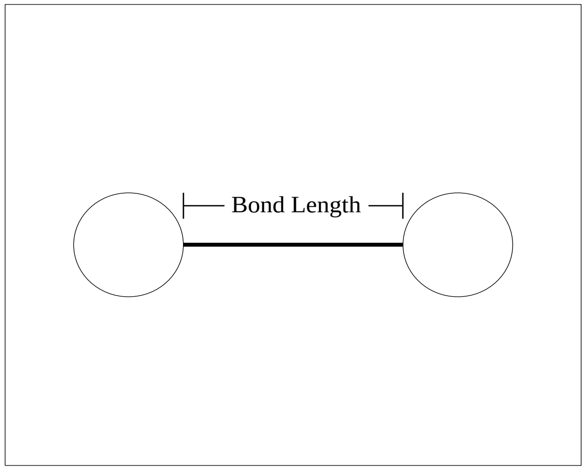
ϕ ψ

ϕ, ψ

ϕ, ψ

	ϕ	ψ
α		
α		
10		
β		
β		

tertiary structure *conformation*



native

°

x_i, y_i, z_i $\leq i \leq n$ n

C_{α_1}

d_{ij} $\overline{x_i - x_j \quad y_i - y_j \quad z_i - z_j}$

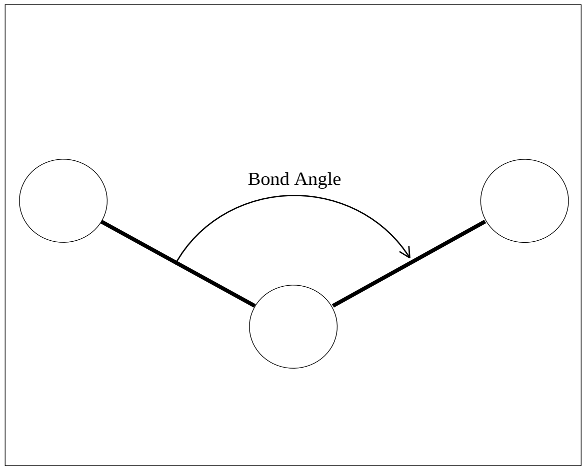
n

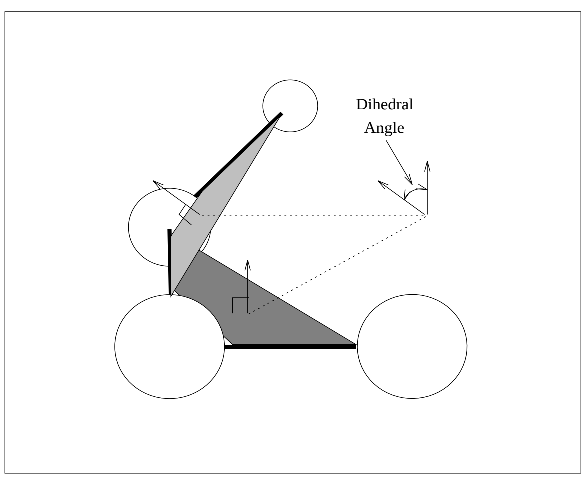
bond length

bond angle *dihedral angle*

$n -$ n

ϕ ψ ω χ_i


Bond Angle


Dihedral
Angle

d

d^N N

$\circ$ $-$ $\circ$

$^{24} \approx . \times {}^{30}$

[9], [12] [15]

$. \times {}^{30}$

	$\approx$
	$\approx$
	$\approx$

A.2 Experimental Tertiary Structure Determination

A.3 Tertiary Structure Prediction (PFP)

Exact

A.3.1 Classical Prediction Methods. *Molecular dynamics*

$^{-14}$ $^{-13}$

$^{-15}$

$^{-9}$

extended-atom representation

energy minimization

Ab initio *Semi-empirical*

Force-field

n

$^{-9}$

		n
ab initio	$\mathcal{O}(n^5)$	
semi-empirical	$\mathcal{O}(n^4) - \mathcal{O}(n^3)$	
force-field	$\mathcal{O}(n^2)$	

$$
\begin{aligned}
E \quad & \sum_{(i,j,k,l)\in\mathcal{D}} \frac{U_{0_{ijkl}}}{} \quad \pm \quad n_{ijkl} \quad {}_{ijkl} \\
& \sum_{(i,j)\in\mathcal{N}} \epsilon_{ij} \quad F_{ij} \quad \frac{r_0}{r_{ij}}^{12} - \; . \; \frac{r_0}{r_{ij}}^{6} \\
& \sum_{(i,j)\in\mathcal{N}} \; . \; \frac{q_i q_j}{D r_{ij}} \\
& \sum_{(i,j)\in\mathcal{H}} \epsilon_{ij} \quad \frac{r_0}{r_{HX}}^{12} - \; . \; \frac{r_0}{r_{HX}}^{10}
\end{aligned}
$$

- $\mathcal{D}$ ω χ
- $\mathcal{N}$
- $\mathcal{H}$
- r_{HX}
- r_{ij} i j
- ${}_{ijkl}$ $i, j, k,$ l
- q_i i
- $U_{0_{ijkl}}$ n_{ijkl} F_{ij} ϵ_{ij} r_0 D

$E_{\mathcal{B}}$ $E_{\mathcal{A}}$ $E_{\mathcal{D}}$ $E_{\mathcal{N}}$ $E_{\mathcal{N}'}$

A.3.2 Other Prediction Methods. *homology*

sequence-structure alignment

$$
\begin{aligned}
E \quad & \sum_{(i,j)\in\mathcal{B}} K_{r_{ij}} \; r_{ij} - r_{eq} \;^{2} \\
& \sum_{(i,j,k)\in\mathcal{A}} K_{\Theta_{ijk}} \quad {}_{ijk} - \quad {}_{eq} \;^{2} \\
& \sum_{(i,j,k,l)\in\mathcal{D}} K_{\Phi_{ijkl}} \qquad n_{ijkl} \;\; {}_{ijkl} - \gamma_{ijkl} \\
& \sum_{(i,j)\in\mathcal{N}} \quad \frac{A_{ij}}{r_{ij}} \;^{12} - \quad \frac{B_{ij}}{r_{ij}} \;^{6} \qquad \frac{q_i q_j}{\pi \varepsilon r_{ij}} \\
- & \sum_{(i,j)\in\mathcal{N}'} \quad \frac{A_{ij}}{r_{ij}} \;^{12} - \quad \frac{B_{ij}}{r_{ij}} \;^{6} \qquad \frac{q_i q_j}{\pi \varepsilon r_{ij}}
\end{aligned}
$$

- $\mathcal{B}$
- $\mathcal{A}$
- $\mathcal{D}$
- $\mathcal{N}$
- $\mathcal{N}'$
- r_{ij} i j
- ${}_{ijk}$ $i, j,$ k
- ${}_{ijkl}$ $i, j, k,$ l
- q_i i
- $K_{r_{ij}}$ r_{eq} $K_{\Theta_{ijk}}$ ${}_{eq}$ $K_{\Phi_{ijkl}}$ γ_{ijkl} A_{ij} B_{ij} ε

Simplification

Lattice

population *encoded* *operators* *fitness*
function *generation*

always

irregular

good

B.1 Brief History of Evolutionary Algorithms

Evolutionstrategie

V $\langle \mathbf{x}, \sigma \rangle$ $\mathbf{x}$ σ

t

$$\mathbf{x}^{t+1} \quad \mathbf{x}^{t} \quad N \quad , \sigma$$

σ

B.2 Origins of Genetic Algorithms

[1]

"Adaptation in Natural and Artificial Systems"

Schema Theorem *Fundamental Theorem of Genetic Algorithms*

B.3 Simple Genetic Algorithm (SGA)

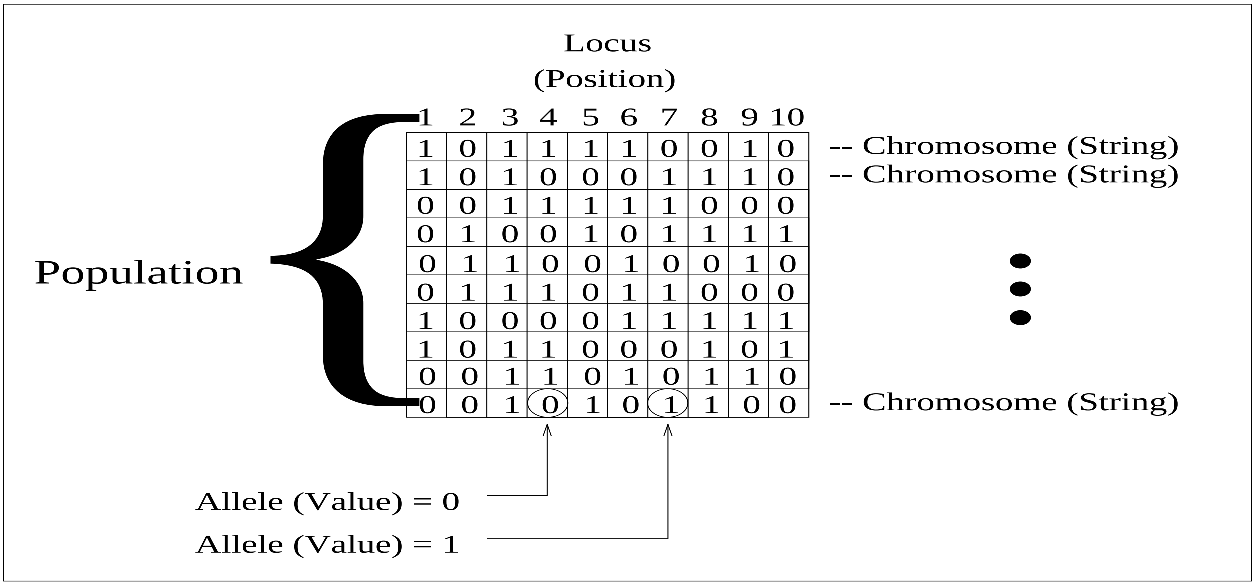
string *chromosome* *genes*

locus *allele*

[1] The term **phenotype** refers to the traits expressed by an individual, in this case the value returned by a function. Contrast this with **genotype** which refers to the traits that define the individual, for example the parameters of the function

population

evolves



x_{min} x_{max}

x_{min} x_{max}

B.3.1 Simple Genetic Algorithm Operators.

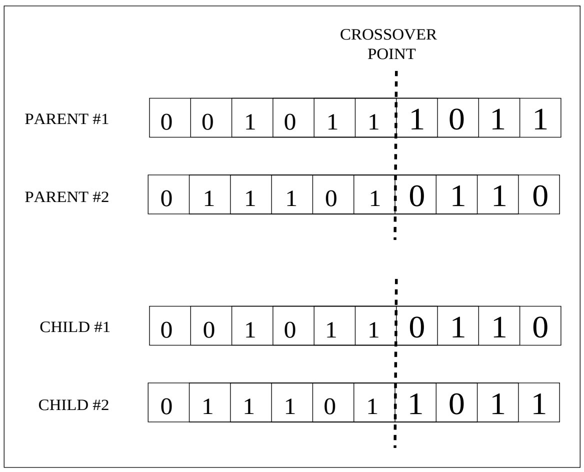
selection *crossover* *mutation*

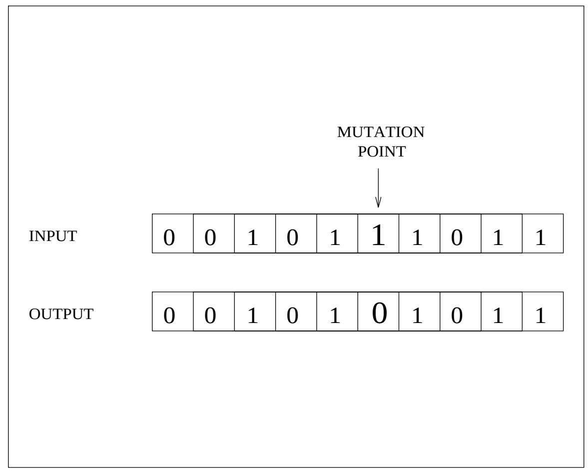
selected *Crossover*

converge

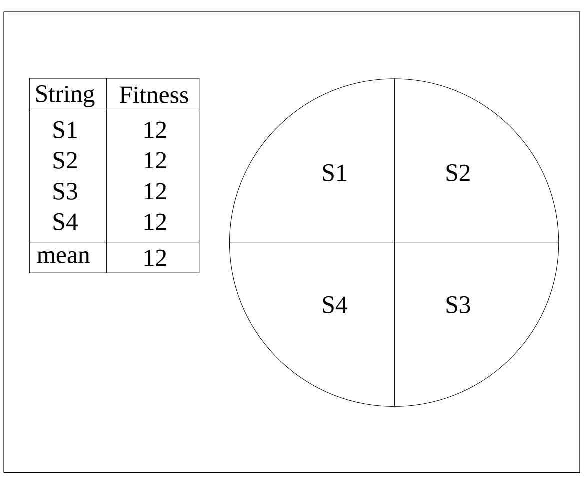
Mutation

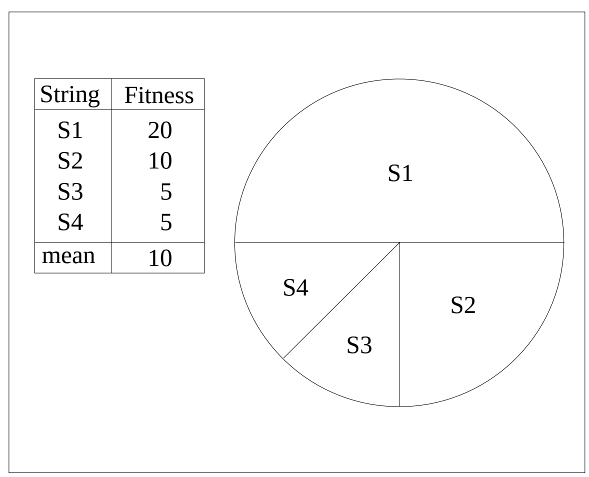
single-point crossover *bitwise mutation*



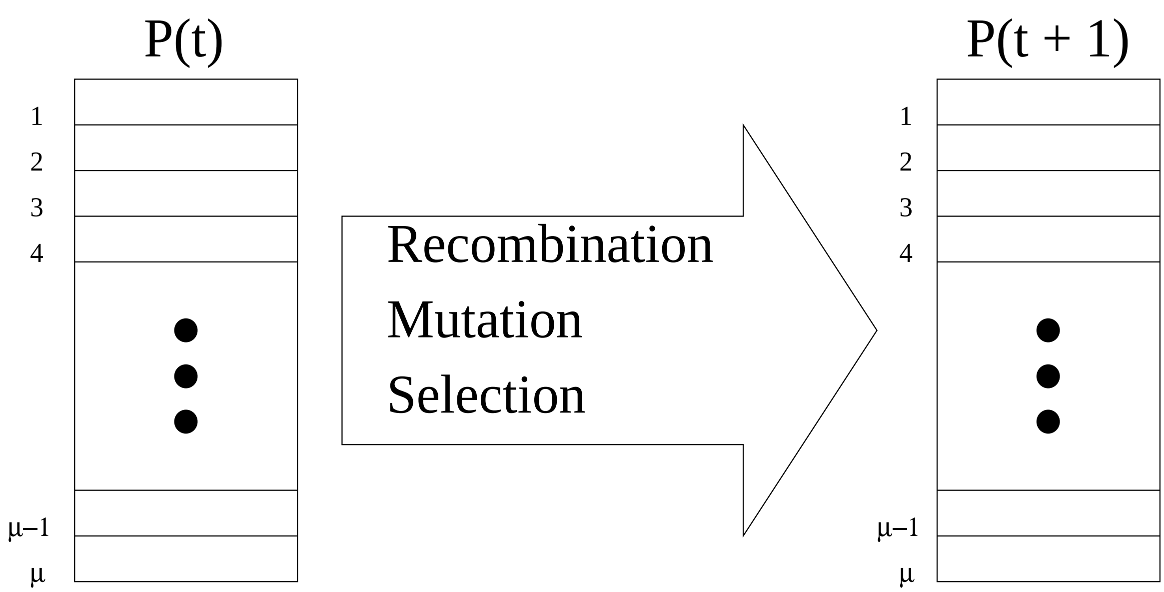


proportional *roulette-wheel*





S S S S *Rank-based* *tournament* *elitist*



B.3.2 *Simple Genetic Algorithm Parameters.*

on-line performance

B.3.3 *Mathematical Theory of How (Why) Simple GAs Work.* *Schemata*

don't care $\{$ $\}$ $\{$ $\}$ *defining length* δH *order* $o H$ H

$\delta *$ $-$ $o *$ $\delta * *$ $-$ $o * *$

$$m\ H,t \quad \geq m\ H,t\ \cdot \frac{f\ H}{f} \quad - p_c \frac{\delta\ H}{l -} - o\ H\ p_m \quad ,$$

H

$m\ H,t$

$m\ H,t$

H $\frac{f(H)}{\bar{f}}$

H $p_c \frac{\delta(H)}{l-1}$

H $o\ H\ p_m$

l *implicit parallelism*

B.3.3.1 Complexity Analysis.

$\mathcal{O}\ nl$ n l

$\mathcal{O}\ nl$

$\mathcal{O}\ nl$

B.4 Messy Genetic Algorithm (mGA)

building block

Deception

linkage

under-specified *over-specified*

competitive template

B.4.1 Messy Genetic Algorithm Operators.

thresholding

cut-and-splice

cut *splice*

primordial *juxtapositional* *partially enumerative initialization*

(PEI) k

k $\binom{l}{k}$

B.4.2 Messy Genetic Algorithm Parameters.

B.4.3 Mathematical Theory of How (Why) Messy GAs Work.

normalized expected defining length $\frac{\langle\delta\rangle}{l+1}$ k

$$\frac{\langle\delta\rangle}{l} \quad \frac{k -}{k}$$

B.4.3.1 Complexity Analysis.

$\mathcal{O}\ l^k$

$\mathcal{O}\ l$ l

n *much*

B.5 Fast Messy Genetic Algorithm (fmGA)

B.5.1 Fast Messy Genetic Algorithm Operators.

probabilistically complete initialization

$l - k$

k

B.5.2 Fast Messy Genetic Algorithm Parameters.

B.5.3 Mathematical Theory of How (Why) Fast Messy GAs Work.

B.5.3.1 Complexity Analysis.

$\mathcal{O}\ l \quad l$

$\mathcal{O}\ l \quad l \quad \mathcal{O}\ l^k$

$\mathcal{O}\ nl$

Appendix C. Background on Parallel Computing

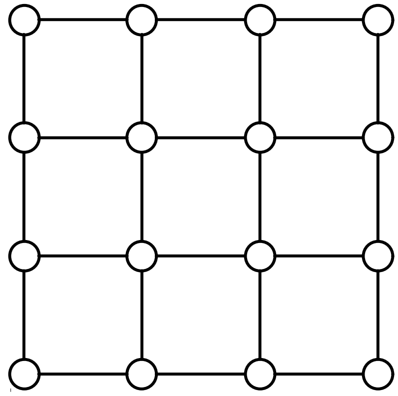
C.1 Parallel Architectures

node

granularity

fine grained

coarse grained



$\times$

shared memory

distributed memory

message passing

interconnection topology *network*

2-D mesh

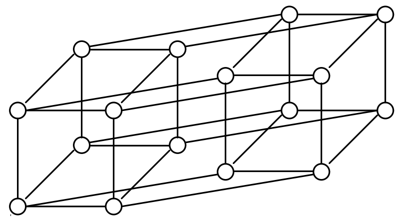
$\times$

hypercube

dimension N ^{N}

N

N



C.2 Parallel Algorithms.

in in.something
something in

in.something

>
/dev/null

psga_param
psga_default psga_param
psga_default
in

mpirun
-np
-sz

-
-

in.2.10.24
& >
/dev/null

```
     Experiments = 1
    Total Trials = 500
 Population Size = 20
Structure Length = 240
  Crossover Rate = 0.65
   Mutation Rate = 0.005
  Generation Gap = 1.0
  Scaling Window = 1
 Report Interval = 1
Structures Saved = 1
Max Gens w/o Eval = 10
   Dump Interval = 0
     Dumps Saved = 0
         Options = ycel
 Number of Peaks = 1.0
Minimization Prob = 1.0
Replacement Prob = 1.0
     Random Seed = 987654321
        Rank Min = 1.5
```

Appendix E. Genocop-III

E.1 Algorithm

Procedure Genocop III
begin
  $t \leftarrow$ t
  P_s t
  P_r t
  P_s t
  P_r t
  while not **do**
  begin
    $t \leftarrow t$
    P_s t P_s $t -$
    P_s t
    P_s t
    if t *mod* k **then**
    begin
      P_r t
      P_r t P_r $t -$
      P_r t
    end
  end
end

procedure evaluate P_s t

begin

$\vec{s} \in P_s$ t **do**

if $\vec{s} \in \mathcal{F}$

then $\vec{s}$ f $\vec{s}$ **else**

begin

$\vec{r} \in P_r$ t

$\vec{z} \in \mathcal{F}$

$\vec{s}$ f $\vec{z}$

if f $\vec{r}$ $>$ f $\vec{z}$ **then** $\vec{r}$ $\vec{z}$ P_r

$\vec{s}$ $\vec{z}$ P_s p_r

end

end

P_s

E.2 Input Parameters

[1]

algorithm domain

application domain

fixed

adaptive

`genocop.h`

-
-
-
-
-

[1] Here the phrase "Domain Constraints" is used in a context more limited than normally used in compute science, specifically, the allowable range of specified variables. If a domain constraint is not defined for a variable, it defaults to the architecture dependent range for $\boldsymbol{R}$

		$N \cup$
		$N \cup$
		$N \cup$
		$N \cup$
		$N \cup$
		N
		N
		N
		N
		N
		N
		N
		$\boldsymbol{R}$
		$\le iSeed \;\le$
		$\le iSeed \;\le$

E.3 Operators

E.3.1 Whole arithmetical crossover. $\vec{z}$ $a\vec{x}$ $-a$ $\vec{y}$

$\vec{x}$ $\vec{y}$ $\leq a \leq$

E.3.2 Simple arithmetical crossover.

single point crossover

E.3.3 Whole uniform mutation. $\vec{x}$ $x_1, ..., x_k, ...x_n$

$\vec{x}'$ $x_1, ..., x_k', ...x_n$ x_k'

k

E.3.4 Boundary mutation. x_k' *left* k *right* k

E.3.5 Non-uniform mutation.

$$x_k^{t+1} \quad \begin{matrix} x_k^t \quad \triangle \; t, r \;\; k \;\; - x_k \\ x_k^t - \triangle \; t, x_k - l \;\; k \end{matrix}$$

k $, \ldots, n$ $\triangle \; t, y$ $, y$

$\triangle \; t, y$ t t

t

$$\triangle \;\; t, y \qquad y \cdot r \cdot \qquad - \frac{t}{T} \;\; ^{b},$$

r .. T b

E.3.6 Whole non-uniform mutation.

E.3.7 Heuristic crossover. *heuristic crossover*

$\vec{z}$

$\vec{x}$ $\vec{y}$

$$\vec{z} \quad r \cdot \quad \vec{x} - \vec{y} \quad \vec{x}$$

r $\vec{x}$ $\vec{y}$ $f \ \vec{x} \ \geq f \ \vec{y}$

$f \ \vec{x} \ \leq f \ \vec{y}$

E.3.8 Gaussian mutation. *Gaussian mutation*

$\vec{x} \quad \langle x_1, \ldots, x_n \rangle$

$$\vec{x}^{t+1} \quad \vec{x}^{t} \quad N \quad , \vec{\sigma}$$

$N \quad , \vec{\sigma}$

$\vec{\sigma}$

$\vec{\sigma}$

E.3.9 Pool recombination operator.

E.3.10 Scatter search operator.

$k >$ J

$\vec{b}$ $\vec{w}$ $\vec{c}$

$$\vec{c} \quad \sum_{\vec{x}_i \in J - \vec{w}} \vec{x}_i / \ k -$$

$\vec{y}$

$$\vec{y} \quad \vec{c} \quad \vec{c} - \vec{w}$$

orgy

F.1 Analysis of Variance (ANOVA)

F.1.1 Single Factor Factorial Design .

a

n a

$$y_{ij} \quad \mu \quad \tau_i \quad \epsilon_{ij} \quad \begin{matrix} i & , & , \ldots, a \\ j & , & , \ldots, n \end{matrix}$$

y_{ij} ij μ *overall mean* τ_i

i *treatment effect* ϵ_{ij}

σ^2

fixed effects model **only**

random samples *random effects model* τ_i

τ_i

a n

N $\sum_{i=1}^{a} n_i$

analysis of variance

$$SS_T \quad \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} - y_{..} \quad ^2$$

SS_T

$$SS_T \quad SS_{Treatments} \quad SS_E$$

$SS_{Treatments}$ SS_E

no difference $H_0 \quad \mu_1 \quad \mu_2 \quad \ldots \quad \mu_a$

$H_0 \quad \tau_1 \quad \tau_2 \quad \ldots \quad \tau_a$

$$F_0 \quad \frac{SS_{Treatment}/\ a-}{SS_E/\ N-a} \quad \frac{MS_{Treatments}}{MS_E}$$

F $a-$ $N-a$ H_0

$$F_0 > F_{\alpha,\ a-1,\ N-a}$$

F_0

$$SS_T \quad \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{ij}^2}{N}$$

$$SS_{Treatments} \quad \sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N}$$

$$SS_E \quad SS_T - SS_{Treatments}$$

y

				F_0
	$SS_{Treatments}$	$a-$	$MS_{Treatments}$	$F_0 \quad \frac{MS_{Treatments}}{MS_E}$
	SS_E	$N-a$	MS_E	
	SS_T	$N-$		

$$y_{i.} \quad \sum_{j=1}^{n} y_{ij}, \quad y_{i.} \quad y_{i.}/n \quad i \quad , \ , \ldots, a$$

$$y_{..} \quad \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \quad y_{..} \quad y_{..}/N$$

F.1.2 Two Factor Factorial Design .

a A b B

$$y_{ijk} \quad \mu \quad \tau_i \quad \beta_j \quad \tau\beta_{ij} \quad \epsilon ijk \quad \begin{matrix} i \quad , \quad , \ldots, a \\ j \quad , \quad , \ldots, b \\ k \quad , \quad , \ldots, n \end{matrix}$$

μ τ_i i A β_j j B $\tau\beta_{ij}$ τ_i β_j ϵijk A

$$H_0 \quad \tau_1 \quad \tau_2 \quad \ldots \quad \tau_a$$
$$H_1 \quad \exists\, \tau_i \; /$$

B

$$H_0 \quad \beta_1 \quad \beta_2 \quad \ldots \quad \beta_b$$
$$H_1 \quad \exists \beta_i \; /$$

A B

$$H_0 \quad \tau\beta_{ij} \quad \forall\; i,\; j$$
$$H_1 \quad \exists\, \tau_i \; /$$

H_0

				F_0
	SS_A	$a -$	$MS_A \quad \frac{SS_A}{a-1}$	$F_0 \quad \frac{MS_A}{MS_E}$
	SS_B	$b -$	$MS_B \quad \frac{SS_B}{b-1}$	$F_0 \quad \frac{MS_B}{MS_E}$
	SS_{AB}	$a - \quad b -$	$MS_{AB} \quad \frac{SS_AB}{(a-1)(b-1)}$	$F_0 \quad \frac{MS_{AB}}{MS_E}$
	SS_E	$ab \;\, n -$	$MS_E \quad \frac{SS_E}{ab(n-1)}$	
	SS_T	$abn -$		

$$SS_T \quad \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}{}^2}{abn}$$

$$SS_A \quad \sum_{i=1}^{a} \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn}$$

$$SS_B \quad \sum_{j=1}^{a} \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{abn}$$

$$SS_{AB} \quad SS_{Subtotals} - SS_A - SS_B$$

$$SS_E \quad SS_T - SS_{Subtotals}$$

$$SS_{Subtotals} \quad \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{abn}$$

$$y_{i..} \quad \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}, \quad y_{i..} \quad y_{i..}/bn \quad i \quad , \; , \ldots, a$$

$$y_{.j.} \quad \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}, \quad y_{.j.} \quad y_{.j.}/an \quad j \quad , \; , \ldots, b$$

$$y_{ij.} \quad \sum_{k=1}^{n} y_{ijk}, \quad y_{ij.} \quad y_{ij.}/n \quad \begin{matrix} i \quad , \; , \ldots, a \\ j \quad , \; , \ldots, b \end{matrix}$$

$$y_{..} \quad \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ij}, \quad y_{...} \quad y_{...}/abn$$

F.2 Kruskal-Wallis H Test.

k

n k R_i

i

Suppose we have k independent samples from k populations. We wish to test the null hypothesis

H_0 : the samples are from identical populations

against the alternative hypothesis

H_1 : the populations are not identical

at the α level of significance.

1. Compute h. *Calculate*

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

2. Accept or reject H_0. *If $h > \chi^2_{k-1,\alpha}$, reject H_0; otherwise accept H_0.*

Probability, Statistics, and Queueing Theory: With Computer Science Applications

Journal of Global Optimization

Evolutionary Algorithms in Theory and Practice

Algorithmics: Theory and Practice

Journal of Computational Chemistry

Proceedings of the Fourth International Conference on Genetic Algorithms

Physics Today

Parallel Program Design: A Foundation

Journal of Global Optimization

Grand Challenges 1993: High Performance Computing and Communications

Protein Folding

Proteins: Structures and Molecular Properties

Handbook of Genetic Algorithms

International Conference on Genetic Algorithms

Proceedings of the Fifth Interantional Conference on Genetic Algorithms

An Analysis of the Behavior of a Class of Genetic Adaptive Systems

IEEE Transactions on Systems, Man and Cybernetics

Genetic Algorithms and their Applications: Proceedings of the Second Internation Conference on Genetic Algorithms

Parallel Genetic Algorithms

The Second Annual Conference on Evolutionary Programming

Task Scheduling in Parallel and Distributed Systems

Proceedings of the Third International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

The Mathematica Journal

International Conference on Genetic Algorithms

Proceedings of the Second IEEE Conference on Evolutionary Computation

The Second Annual Conference on Evolutionary Programming

Artificial Intelligence Through Simulated Evolution

IEEE—ACSSC

Foundations of Genetic Algorithms 2

Computers and Intractibility: A Guide to the Theory of NP-Completeness

Decision Sciences

Genetic Algorithms in Search, Optimization, and Machine Learning

Genetic Algorithms, Noise, and the Sizing of Populations

Proceedings of the Fifth International Conference on Genetic Algorithms

Complex Systems

International Conference on Genetic Algorithms

Complex Systems

Proceedings of the Second International Conference on Genetic Algorithms

Proceedings of the Fifth International Conference on Genetic Algorithms

IEEE Transactions on Systems, Man & Cybernetics

Proceedings of the Third International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

Deception Considered Harmful

Proceedings of International Parallel Processing Symposium

Adaptation in Natural and Artificial Systems

Scientific American

Principles of Biochemistry

Biochemistry

Applied Computing 1997: Proceedings of the 1997 Symposium on Applied Computing

The Origins of Order, Self-Organization and Selection in Evolution

Advances in Parallel Algorithms

Introduction to Parallel Computing: Design and Analysis of Algorithms

Compendium of Parallel Programs for the Intel iPSC Computers

Calculus with Analytic Geometry

Journal of Global Optimization

The Protein Folding Problem and Tertiary Structure Prediction

Arbeitspapiere der GMD 748

Real Time System Design

Introduction to Parallel Computing

Proceedings of the National Academy of Science USA

Theory and Problems of Organic Chemistry

Applied Computing 1996: Proceedings of the 1996 Symposium on Applied Computing

Evolutionary Computation

Proceedings of the 3rd Annual Conference on Evolutionary Programming

Evolutionary Computation Journal

Genetic Algorithms + Data Structures = Evolution Programs

Genetic Algorithms + Data Structures = Evolution Programs

Molecular Structure (Theochem)

Design and Analysis of Experiments

Paralle Computing

Journal of Computational Chemistry

SPIE Proceedings 1

Heuristics

Proceedings of the Third International Conference on Genetic Algorithms

Numerical Recipes in C: The Art of Scientific Computing

Protein Science

Third International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

MPI: The Complete Reference

WEBSTER'S II New Riverside University Dictionary

IEEE-CH

Proceedings of the Fourth International Conference on Genetic Algorithms

COMPUTER

Proceedings of the Fourth International Conference on Genetic Algorithms

Biochemistry

Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms

genitor
International Conference on Genetic Algorithms

PPSN III

Proceedings of the Fourth International Conference on Genetic Algorithms

Foundation of Genetic Algotithms

Vita